LATENT VARIABLE REALISM IN PSYCHOMETRICS

Steven Brian Hood

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of History and Philosophy of Science,
Indiana University
May 2008

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Colin Allen, Ph.D.

_____

Denny Borsboom, Ph. D.

_____

Sander Gliboff, Ph. D.

_____

Amit Hagar, Ph. D.

_____

Mark Kaplan, Ph. D.

May 5, 2008

To my mentors:
Jarrett Leplin,
Dodona Kiziria,

&

To my friend:
Tyson “Six-shooter” Sadleir

ACKNOWLEDGEMENTS

Many persons have left their marks on this dissertation. Above all I owe an incredible debt to my Chair, Colin Allen. I am truly fortunate to have had Colin as an advisor. If I were to footnote every point in this dissertation where Colin contributed insight or offered guidance, the page count would have skyrocketed, making it prohibitively expensive to have the dissertation bound. I am especially appreciative of the encouragement Colin gave me to write the dissertation that I wanted to write. He offered much in the way of pragmatic advice for effectively carrying out the project that follows. I'm thankful for the generous attention that Colin dedicated to reading over drafts of sections, always offering copious comments and suggestions. With Colin's support I secured a research fellowship at Tilburg University's Center for Logic and Philosophy of Science during the spring of 2008. It was in Tilburg where I finished the dissertation. Being in the Netherlands allowed me to work closely with the Psychological Methods group at the University of Amsterdam and the faculty at Tilburg University. In addition to his contributions as Chair, Colin inspires. He is a role model, both professionally and personally, to every student who has the good fortune to work with him.

Sandy Gliboff made many helpful comments on early drafts. He has also supported me since I came to IU in many respects not all related to the production to this dissertation. Though I came from a philosophy background, the inspiration I got from Sandy more than once made me question my intellectual identity: should I be a philosopher or a historian? Eventually I settled on the former. All the better for history, I suppose.

Amit Hagar, who joined the department after I was already well into writing the dissertation, generously agreed to serve on the committee. He was especially instrumental in helping me understand ergodicity, a concept borrowed form statistical thermodynamics and central to the arguments in Chapter 5.

Mark Kaplan offered many helpful suggestions at my proposal defense which have helped to shape the dissertation. Taking his course on decision theory planted the seeds for the arguments in Chapter 3. Given that there's really no decision theoretic concepts occupying center stage in that chapter, Mark's influence may not be obvious, but the idea of discerning one's commitments based on one's actions is an idea that I got from Mark; however, this shouldn't be read as attributing to him any particular views on methodology.

Denny Borsboom's contribution to this project is immeasurable and his mark is evident throughout this dissertation. Denny's work is the *sine qua non* for what I say in the following pages. It was his (2003) paper on the theoretical status of latent variables that inspired this entire dissertation. In reading the rest of his work I found many issues where I thought philosophical perspective could help clear up matters. So if I have anything interesting to say about latent variables or validity, I ride comfortably and proudly on Denny's coattails.

I would not have even had the opportunity to meet Denny Borsboom were it not for the support of the Department of History and Philosophy of Science. In March and November of 2007, I was fortunate to be able to travel to the University of Amsterdam. These trips, made possible by the Sam Westfall Graduate Student Travel Grant, were invaluable to my research. There Denny and other members of the Psychological Methods group were generous with their time and department resources, and they provided valuable comments on my work. I am especially grateful to Conor Dolan and Gideon Mellenbergh who helped me sort out some issues pertaining to measurement models versus structural equation models, reliability, and validity.

With Conor Dolan and Denny Borsboom, I co-authored a paper comparing bifactor and higher order factor models of intelligence data. This paper appears in a slightly altered form as Chapter 4. The ideas for that chapter were the product of many conversations with Conor that took place in the hallways of the fifth floor of the Psychology building at the University of Amsterdam. Conor would stop me on my way to get coffee to discuss the *g* factor and the next thing I knew, one of the departmental bulletin boards became a confirmatory factor model with

tacks representing latent variables and observable indicators. In addition to spawning Chapter 4, these conversations helped to shape the way I think about measurement models and latent variables. I am also grateful to Conor for allowing me to stay at his place during the last month of January 2008 after my housing contract expired. I can honestly say that Conor's influence kept me off the streets.

Benjamin Lovett has been a continual source of support, spending countless hours on the phone with me over the past couple years discussing *g* and IQ. It was while working for Johns Hopkins University's Center for Talented Youth in 2001 that Ben and I began bantering back and forth about the nature of intelligence and Flynn's Effect. Such were the seeds of this project. Years later he sent me Denny Borsboom's paper on the theoretical status of latent variables and a dissertation proposal was born.

If one goes back far enough in the etiology of this project, one will find Jarrett Leplin. Jarrett is the reason I went into philosophy at all. I wrote this dissertation with the aim of not only expressing my own thoughts about psychological measurement, but I also wanted to pay homage to my intellectual benefactor and mentor. It is from him that I learned about realism, epistemology, and culinary sensibility. If there is one sentence in this dissertation or one argument with which he is pleased (and I get a job), then my goal will have been accomplished. Anyone who knows Jarrett or his work will see in this dissertation his influence on me. I will unabashedly admit to borrowing his ideas and tapping his inspiration at every point possible. Any philosopher of science interested in the epistemology of science owes it to himself to read his work.

Also, this dissertation has benefited from conversations I have had with Steve Crowley, Hilmi Demir, Daniel Dennett, Melinda Fagan, John Johnson, Kent van Cleave, Ginette Delandshere, and Grant Goodrich. Jocelyn Holden generously devoted her time to reading drafts of several chapters and in offering her expertise in structural equation modeling and factor analysis. Chapters 1, 2, and 3 especially benefited from her insight and statistical savvy. The

section of Chapter 3 on item response theory was spawned by a late night conversation with Jocelyn at a party. In vino veritas?

Matthew Dunn, my roommate during three years of my working on this project, offered continued patience in discussing my ideas and helping me flesh them out. He remains skeptical of pretty much everything I say. Christopher Martin provided encouragement and emotional support, as well as well-needed distractions.

The departmental staff, Peg Roberts and Becky Wood, kept me on time and aware of deadlines. Without their support I would never have finished this project, for I would have most certainly remained lost in the labyrinth of bureaucracy.

I must also thank my parents Ted Hood, Regina Goree, and Robert Goree for constantly reminding me (and I mean *constantly*) that "all you have to do is write the dissertation and you'll be done." In all seriousness however, they've been a wonderful source of moral support even if they still have no clue exactly what I do.

And Sebastian, my companion and dear friend, I live to see your little nub wag. Your love is unconditional. You provided me with many attractive opportunities not to work on this dissertation. For a canine, you did a wonderful job of keeping me human. He's a good stinky pups.

ABSTRACT

LATENT VARIABLE REALISM IN PSYCHOMETRICS

This dissertation concerns the theoretical status of latent variables in psychometrics and the philosophical foundations of psychometrics. I work toward the construction of a philosophical framework for psychometrics by examining and refining fundamental psychometric concepts such as *validity,* and by proposing a theoretical interpretation of latent variables. I mine psychometric methods for tacit philosophical commitments, make them explicit, and evaluate them. With its philosophical presuppositions made explicit, I then articulate what are realistic epistemic aspirations for psychometrics

Latent variables cannot be measured directly; they must be inferred from observable variables phenomena. For example, variability in scores on psychometrics tests, are "explained" by positing an unobservable source of the observed variability, i.e., a latent variable. In psychometrics, many central theoretical constructs have their provenance in latent structure analysis. I focus on a particular latent variable, *g*, the general factor of intelligence.

At the most general level, this dissertation will address the following questions:

1. Does psychometric practice require regarding abilities as real entities?
2. Is epistemic realism regarding *g* tenable?

I argue that the answer to (1) is "yes" irrespective of any one psychometrician's professed philosophical commitments. However, unlike the critics of psychometrics, I argue that this commitment is innocuous and productive.

The answer to (2) depends on what it would mean to be a realist about *g*. I argue that *g* is not a causally efficacious attribute of individuals. It makes sense only as a relation *between* individuals. Statistical models of *g* do not contain information about causal processes within

individuals. While it may be wrong to conceive of *g* as causally efficacious, models of *g* do constrain theories of mental ability at the level of individuals.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Figures

Tables

CHAPTER ONE

LATENT VARIABLES, THE GENERAL FACTOR OF INTELLIGENCE, AND TWO CRITICISMS

…if we take the universe of "fitting," countless coats "fit" backs, and countless boots "fit" feet, on which they are not practically fitted; countless stones "fit" gaps in walls into which no one seeks to fit them actually. In the same way countless opinions "fit" realities, and countless truths are valid, tho' no thinker ever thinks them.

–William James

1. Introduction: Variables Latent, Variables Manifest

Latent variables are ubiquitous in the social and behavioral sciences. Some would even say that they are an indispensable part of social and psychological research (Sobel, 1994). At a general level we may distinguish between variables that are *manifest* and variables that are *latent*. Manifest variables are distinguished by being observed. Suppose we set out to measure the lengths of various objects. With a meter stick in hand, we get to work: the length of the swimming pool is 100 meters, the length of my hotdog is a little under a third of a meter, the

ceiling is three meters from the floor in a room, etc. In this case, length is a manifest variable. It is an observed property of the objects we measure. Contrast the case of manifest variables with the following: we want to find out someone's or a demographic group's socioeconomic status (SES). However, instead of setting out with an instrument that measures SES directly, we pass out questionnaires asking our subjects to report their gross annual income, level of education and occupations of both the subject and his parents. Based on the values for each of those manifest variables, we locate our subjects on the SES index. SES is not observed directly, it is a composite score based on values for manifest variables.

Some latent variables, such as general intelligence, are essentially latent; others are latent as a matter of methodological consequence or convention. Variables that would be manifest, were they observed, would be considered latent if they figure in as a latent variable in a latent variable model. A latent variable model specifies the relationship (either quantitative or qualitative) between various observable indicators (manifest variables) and a latent variable. Hence, length, though a manifest variable in the example considered above, could be a latent variable depending on one's measurement methods. For example, we may want to assess the lengths (i.e., height) of adolescents, in a town where there are no meter sticks, rulers, or other devices for measuring length. Instead we may record their weights and shoe sizes and use those values to infer values for height. On the basis of those measures we should be able to predict with good accuracy the length of the adolescents since the values for the manifest variables are known to correlate highly with height in adolescents.

What makes height a latent variable in this example is that it is not observed; rather it is inferred on the basis of observed and known indicators of height. Hence, whether a variable is latent depends on the method used to ascertain its value in a particular instance. Note that whether a variable is latent or not does not depend on a particular measurement method such as factor analysis, structural equation modeling, or any other statistical method. There are plenty of readily imaginable cases of latent variables that do not require nor do they have their provenance in

statistical methods. Statistical methods are just members of a family of methods for acquiring information about the world. Certainly, using some statistical methods will be sufficient for ensuring that the variable under investigation is latent, but there is no essential feature of latent variables in general that depends on statistical methods.

From what has been said thus far it should be clear that latent variables form a heterogeneous bunch, united only by the fact that they are not manifest. A dissertation about latent variables in general would be a dissertation about unobservability. I will not be concerned with whether there is a meaningful metaphysical distinction between being observable and being unobservable. Also, It is not clear that such an investigation would be all that enlightening to the practicing scientist since science itself is made up of a heterogeneous mish-mash of methods and aims. Consequently, I will restrict my investigation of latent variables to the social and behavioral sciences and the inference problems introduced by positing latent variables, namely how we go from latent variables to quantities in nature. But even this is to cast the net a bit widely. Specifically I will be concerned with latent variables in psychometrics, a branch of psychology devoted to the investigation of psychological traits and the structure of individual and group differences in psychological traits. Educational assessment, personality assessment, psychological measurement, and the construction of psychological tests all fall within the purview of psychometrics. I will devote considerable attention to the general factor of intelligence, i.e., the *g*-factor. The reasons for focusing on *g* as opposed to any other psychometric factor are manifold, though much of what follows in this dissertation can be generalized to latent variables in general. When necessary, I will consider latent variables in other disciplines such as socioeconomics or physics.

2. Ways to Think about Latent Variables

Though latent variables may be invoked to refer to unobservable objects, such as electrons or quarks, as they factor in psychometrics and the social sciences, latent variables are typically taken

to refer to properties. I will assume that properties, or at least property instances, have causal powers, either active or dispositional. Thus, if a latent variable successfully refers to a property or property instance, then its referent has causal powers (i.e., it is causally efficacious). This rules out the possibility of epiphenomenal latent variables, and this might seem a bit suspect given that I am dealing with latent variables that purportedly refer to mental properties. Psychometrics seems to presuppose that epiphenomenalism is false, since psychological attributes, the referents, of latent variables are alleged to be causally efficacious (actively or dispositionally) if they exist at all.[1] For my purposes, I assume that epiphenomenalism is either false or a contingent thesis falsifiable by psychometric data.

2.1 Syntax

Let us step back a moment from mental matters and the interpretation of latent variables in psychometrics. I have discussed in a very general way what latent variables are, but in order to go any further, more detail is needed. I will be taking the terms 'latent variable' and 'latent factor' to be synonymous, switching between the two depending upon the context. If the context is factor analysis, I will often use the latter term. If the context is other than factor analysis such as item response theory, I will use the former term. In both contexts we are dealing with latent variables, we just get at them in different ways. The distinction is made to indicate differences in methods, not their products. In what follows in this section, examples will be drawn exclusively from psychometrics; however, the mathematical core of latent variable theory is general. It is not peculiar to psychological measurement.

[1] However, some psychometricians believe that a psychological attribute, A, is causally efficacious only if there is actual variability in A. That is, a disposition to effect change is not strong enough. This variability in position on A is manifested in test behavior. See Borsboom (2005) and Holland (1986). One interesting consequence of this position is that general intelligence, the purported referent of the *g*-factor, is not causally efficacious since it exhibits no variability within individuals, i.e., there is no intraindividual variability. The severity of this consideration for theories of intelligence that take general intelligence to be a central theoretical posit and whether interindividual variation is sufficient to save general intelligence *qua* causally efficacious psychological attribute are questions that will be addressed in chapter 4.

Factor analysis is one statistical procedure for discovering latent variables (exploratory factor analysis) and confirming latent variable models (confirmatory factor analysis). In factor analysis, the manifest and latent variables are metrical, i.e., they take real number values (either continuous or discrete). This may be contrasted with latent trait analysis (e.g., IRT, which I discuss in chapter 3) where the latent variable is metrical, but the manifest variables are categorical, e.g., 'yes' or 'no', 'correct' or 'incorrect', or '0' or '1'. The utility of factor analysis is manifold. First, factor analysis is a data reduction technique. Suppose you have an m×m correlation matrix. The correlated items may be performance on psychometric tests or what have you. The larger m is, the greater the number of correlations in the matrix and also the more unwieldy the matrix becomes. Sometimes it might be useful to express the information contained in the correlation matrix with a smaller number of variables. For example, a scientist may find it more economical and cognitively tractable to deal with a 5×20 factor matrix expressing the relationship between the manifest variables and a compendious set of latent factors, rather than a 20×20 correlation matrix. To get at an understanding of factor analysis, I'll introduce an example of a factor analysis of a hypothetical correlation matrix.

Consider the following 8×8 correlation matrix from Jensen (1998, 80):

	V1	V2	V3	V4	V5	V6	V7	V8
V2	5600							
V3	4800	4200						
V4	4032	3528	3024					
V5	3456	3024	2592	4200				
V6	2880	2520	2160	3500	3000			
V7	3024	2646	2268	2352	2016	1680		
V8	2520	2205	1890	1960	1680	1400	3000	
V9	2016	1764	1512	1568	1344	1120	2400	2000

Table 1.1: Hypothetical correlation matrix.

As the number of variables increases, so does the utility of being able to represent the information in terms of a few latent factors. Some of the variables correlate more strongly with each other

than with other tests. For example, V1, V2, and V3 are more strongly mutually correlated than they are with other variables. The correlations between variables can be expressed more economically in terms of a correlation with a latent factor. Factor analysis enables us to transform the correlation matrix above into a *factor matrix* expressing the correlation between each test and an "underlying" factor. The number of factors can be as many as the number of variables (though this would simply reproduce the original matrix). Typically, for a factor to be treated as significant it must have an eigenvalue greater than one; this information is represented often represented in a "screeplot." However, the number of factors may be stipulated in the process of doing a factor analysis. The correlation between a test and a factor is that test's *factor loading*. The portion of the variance that is not captured by common factors is called the test's *uniqueness* (u^2) and is the sum of error variance and the true score variance that is unique to that test, i.e., its *specificity*. It is what is left over when we account for the contribution of common factors. The portion of test score variance that is due to common factors is known as the test's *communality* (h^2). Communality and uniqueness are not going to be important for my purposes, but I provide them in the interest of being thorough. Table 2 below reveals three first-order factors (F1, F2, and F3) that account for intercorrelations in the nine test variables, and the correlation between the three primary factors is accounted for by a second-order factor, *g*. The following factor matrix illustrates this point:

Variable	2nd order *g*	1st order F1	1st order F2	1st order F3	Communality h^2	Uniqueness u^2
V1	.72	.3487	0	0	.64	.36
V2	.63	.3051	0	0	.49	.51
V3	.54	.2615	0	0	.36	.64
V4	.56	0	.42	0	.49	.51
V5	.48	0	.36	0	.36	.64
V6	.40	0	.30	0	.25	.75
V7	.42	0	0	.4284	.36	.64
V8	.35	0	0	.3570	.25	.75
V9	.28	0	0	.2856	.16	.84
Variance	2.2882	.283	.396	.3925	3.36	5.64
%Variance	25.42	3.15	4.40	4.36	37.33	62.67

Table 1.2. Factor matrix for hypothetical correlation matrix in Table 1.

To reconstruct the correlation between any two manifest variables, sum the product of the variables' *g* loading and the product of the variables' loading on the first order factors. Note that at this point no interpretation of these factors or correlations has been made. I have remained neutral, portraying factor analysis as a data-reduction technique and factors of expressions of variability. To interpret these factors as having some connection to cognitive ability takes us beyond the data and into the realm of semantics.

2.2 Semantics and Ambiguity in *g*

In psychometrics, latent factors are sometimes interpreted as conveying some information about cognitive ability or personality. Matters are complicated by the fact that not all latent variables are similarly interpreted; some seem to lend themselves to a realist interpretation more than others. SES, for example, is typically not interpreted as something real, existing independent of its measurement. Intelligence and extroversion, however, are generally construed to be features of humans (if not non-human animals as well). Some latent variables are generated by variability between persons (interindividual variation) and others are generated by variability within person (intraindividual variability). The robustness of *g* across different factor analytic techniques, biological correlates with *g*, and the apparent impossibility of constructing a test of cognitive

ability that does not load on *g* have been taken by proponents of *g*-factor theories of intelligence as evidence that there is a general mental ability underlying all cognitive tasks (or at least those sampled by intelligence tests). This is a bit rough since not all intelligence theorists who take *g* to be a requisite *explanandum* for an acceptable theory of intelligence interpret *g* the same. One source of the heterogeneity in interpretations of *g* is confusion over what *g* is. It is not rare to find prominent intelligence researchers conflating distinct concepts under the name '*g*'. This ambiguity in *g* does not stem from the nature of *g* or factor analysis itself, such as the fact that there are multiple factor solutions for any correlation matrix that admits of a general factor. Rather it stems from a failure to be careful and avoid running distinct statistical concepts together. As a cautionary tale to the reader I will now turn to a discussion of how various prominent researchers have fallen prey to such confusions.

The *g*-factor is a general factor that accounts for the variance in performance on a large battery of mental ability tests. The greater the portion of variance accounted for by the *g*-factor on some test, the greater the *g*-loading of that test. For any three tests, X1, X2, and X3, if X1 is highly loaded on *g* and performance on X2 is more highly correlated with performance on X1 than performance on X3, then X2 is more highly *g*-loaded than X3. Also, the higher the correlations between performance on a battery of tests, the greater the portion of variance for which the *g*-factor will account (to the extent that any first order factors are also highly correlated). Conversely, if performance on a battery of tests is uncorrelated or weakly correlated, the *g*-factor will account for no or little variance.

In this section I will expose a pernicious ambiguity in the way psychologists use '*g*'. The four different notions that are sometimes run together under the term '*g*' are

1. *g*-factor: the most general and latent statistical factor that accounts for some portion of the variance in a correlation matrix,
2. *g*-score: the weighted sum of an individual's scores on variables that comprise the *g*-factor; i.e., one's position on the latent variable, the *g*-factor,

3. general mental ability: the trait or attribute measured by reliable and valid tests of mental ability which load heavily on the *g*-factor; the purported latent cause of variability in between-subject scores on tests of mental ability,
4. *g*-loading: the correlation between a variable indicating performance (i.e., a variable in a matrix of correlations) and the *g*-factor.

As I will show, running these four related concepts together can lead to serious confusion and paradoxical results.

2.21 Ambiguating *g*

2.211 Flynn's Effect

Flynn's Effect is the well-documented, worldwide steady increase in average IQ (Flynn 1984, 1987, 1999). IQ gains are, on average, 3 points per decade since 1932 (Neisser, 1998 p. 13). Opponents of the centrality of the *g*-factor object that if IQ tests measure mental ability (i.e., *g* in the third sense above) and IQ has been increasing, then so must mental ability. The force of the objection comes from the fact that performance on highly *g*-loaded IQ tests is highly correlated with academic and occupational achievement, but IQ gains have not been accompanied by corresponding gains in academic and occupational achievement (Deary 2001, Flynn 1999), which is a counter-intuitive result given that academic and occupational achievement are correlated with mental ability.[2]

A popular response to this objection to *g*-factor theories of intelligence (Miele, 1999; Rushton, 1999) is to claim that the IQ gains are hollow in the sense that the gains reflect improved performance on just the non-*g*-loaded sections of the IQ tests. The rationale behind this suggestion is that if the gains in IQ can be accounted for by performance on those sections or items of the test that are not *g*-loaded, then performance is increasing on those sections or items

[2] The claim that IQ gains have not been accompanied by corresponding gains in academic and occupational achievement has not gone uncontested. Robert Howard, for example, has argued that there are such corresponding gains.

of IQ tests that are not measuring mental ability. This response is an instance of a general strategy for countering Flynn's Effect—to acknowledge that there are IQ-gains, but to deny that there are corresponding gains in general mental ability. This response may seem *ad hoc*, especially in light of the fact that IQ-gains are most pronounced on Raven's Progressive Matrices, the psychometric test said to be the purest measure of general mental ability, i.e., the Raven's measures general mental ability and little else.[3]

There is another response that follows the aforementioned general strategy, and though it may avoid charges of arbitrariness or *ad hoc*-ness, the response is marred by equivocation. The response typically goes as follows: if the variance in performance on psychometric tests (or the correlations between performance on the tests) has remained constant, then so will *g*. Therefore, IQ gains need not accompany gains in g, and since there are no gains in general mental ability, we should expect no gains in achievement. This response equivocates because '*g*' in its first occurrence only makes sense when interpreted as meaning the *g*-factor (a between-subject statistic), whereas in the second occurrence it is intended in the sense of general mental ability (a within-subject phenomenon).[4]

2.212 Psycho-educational Assessment

The most dramatic instance of *g*-ambiguity comes from research in psycho-educational research, particularly research on learning disabilities in high-IQ children. The classic operational definition of learning disabilities is a discrepancy between IQ and achievement (Kavale & Forness, 1995). In the field of educational psychology there has been a growing concern over

[3] The Raven's Progressive Matrices test of cognitive ability is a non-linguistic test involving pattern and picture completion. Test takers are presented with progressively more difficult sequences of pictures that form a pattern and are required to specify, based on the pictures given, what the next picture should be. Raven's Progressive Matrices, often referred to as simply "the Ravens" is considered to provide the purest measure of *g*; it is more highly loaded on *g* than any other test of mental ability.

[4] This, by no means, exhausts the possible or extant responses to Flynn's Effect. For example, Colin Allen has suggested in personal correspondence that academic and occupational achievements are assessed relative to competitors, and if all are going up then it is relative *g* levels, not absolute *g* levels, that will predict (relative) achievement.

misdiagnoses of learning abilities in gifted children. Under the classical operational definition of a learning disability, gifted children are at an increased risk for being diagnosed with learning disabilities—a counterintuitive result. If mental abilities are more differentiated at the high end of the ability spectrum, then ability and achievement are also more like to be more discrepant, leading to increased learning disability diagnosis rates. There are reasons to think that gifted children are not at an increased risk for learning disabilities, but that they are at an increased risk for misdiagnoses of learning disabilities (Lovett, B.J., & Lewandowski, L.J., 2006).

That mental abilities show greater differentiation at higher levels, i.e., they are less highly correlated, is a well-confirmed result (Spearman, 1927; Detterman & Daniel, 1989; Deary & Pagliari, 1991; Detterman, 1991; Neisser, 1999). For this and other reasons, researchers have questioned the legitimacy of learning disabilities as IQ/achievement discrepancy. However, for my purposes, the relevant factor is ability-differentiation. Ability-differentiation in high IQ children entails that IQ tests have lower *g*-loadings for that population. If *g*-loading is conflated with general mental ability, the absurd conclusion follows that gifted children have a lower general mental ability than the rest of the population. Indeed, some psychologists have made this inference on the basis of such this confusion. For example, Ulric Neisser (1999, p. 131) writes:

> Generally speaking, *g* accounts for less and less of the variance as one moves to individuals with higher and higher scores. This means that more intelligent people have more diverse sets of specific abilities; those in the lower range have not developed those abilities and must use what little they have for virtually every test. Less intelligent people have relatively more *g*!

The first occurrence of '*g*' in the passage quoted above refers to the *g*-factor; however, the second occurrence of '*g*' is intended to refer to general mental ability. Detterman (1991) gestures toward this inference (though he does not make it) in an article titled "Reply to Deary and Pagliari: Is *g* Intelligence or Stupidity?" In this article, Detterman points out that Spearman may have been committed to the paradoxical result:

> [Spearman] thought that it was *g* that produced the correlations among tests, and that people differed in the *g* they had. Logically, then, groups with the highest correlations among tests should have the largest amount of *g*. Because, in both data reported by Spearman and in my data, the low-IQ groups had the highest correlations among tests, they also must have the largest amount of *g*. In other words, *g* correlates negatively with intelligence, so *g* must be stupidity, (p. 254).

In this passage, the first and second occurrences of '*g*' refer to general mental ability, the fourth occurrence refers to the *g*-loading of the tests administered, and the third and fifth occurrence refers to general mental ability. Conflating these different concepts leads to the paradoxical result that general mental ability is stupidity. Notice also that the aforementioned paradoxical result entails that IQ tests, purported measures of mental ability, do not measure mental ability, at least not in any meaningful way.

2.22 Disambiguating *g*

Clearly the source of the confusion in the previous two examples is the conflation of levels of mental ability with percent of variance accounted for by *g*. A person's level of general mental ability, i.e., what is represented by a person's *g*-score, need not be commensurate with the amount of variance that the *g*-factor captures in a collection of scores on mental ability tests. The percentage of variance accounted for is not even a meaningful statistic for an individual. *g*-loadings are not *g*-scores or *g* as ability.

A related problem arises when we recognize that the *g*-factor is a population statistic that arises when a diverse battery of mental ability tests are administered to a large population. The *g*-factor is best thought of as population-level statistic that expresses how much of the positive manifold in a correlation matrix can be captured by a single, general factor. The claim that the *g*-factor is an indicator of general mental ability in individuals (to a greater or lesser extent) is logically independent from its statistical nature. Once this inference from the presence of a *g*-factor to the presence of a biological or psychological construct construed as general mental ability has been made, one may then inquire as to what an individual's *g*-score is. How to

calculate an individual *g*-score is somewhat problematic, for it is unknown what the distribution of general mental ability is in the population. By convention (though not without rationale), IQ is normally distributed. Assigning *g*-scores to individuals also will require the assumption that the distribution of scores in the population fit some pre-specified shape. Also by convention (though, again, not without rationale), *g*-scores are likewise assumed to be normally distributed in the population. Since general mental ability is latent, any assignment of a *g*-score is an *estimation* of one's true *g*-score, necessarily. The precision of the estimated *g*-score can be tested using standard statistical techniques. Those tests that are purer measures of general mental ability provide the best *g*-indices (when administered in concert), e.g., Raven's Progressive Matrices and Kaufman Brief Intelligence Test (Beaujean, 2002).

3. Dismantling the Gouldian Preemption of Psychometrics

To some this project may seem otiose in light of Stephen Jay Gould's eloquent, vituperative, and polemical *The Mismeasure of Man*. Originally published in 1981 and revised in 1996 in response to Herrnstein and Murray's controversial *The Bell Curve* (1994), *The Mismeasure of Man* is considered by some to be the death knell of psychometrics. This is an exaggeration of *Mismeasure's* achievement. The psychometric community wasted no time responding to Gould's objections, so I will have little new to offer to this well-worn topic. The purpose of this discussion is to convince the reader that Gould did not have the last word on psychometrics, though the persistence and continued fruitful research in the field should be ample evidence that Gould did not offer compelling objections to the psychometric study of cognitive ability. I will, however, shed new light on the role of reification in the history of psychometrics, and this I take to be my novel contribution to the debate. In later chapters I will return to Gould and see if there is a way of recasting his arguments so that they engage with contemporary psychometric. For now I will limit my discussion to the arguments that appear in *Mismeasure*. The goal of this section is to convince the reader that Gould's *Mismeasure* is just that—a mismeasure of current psychometric

research. I will not concern myself with the issue of whether Jensen inferred between-group heritability of IQ on the basis of within-group heritability of IQ, though this is one of the two main charges levied against Jensen. Neven Sesardic (2000) has shown, persuasively in my estimation, that such accusations are ill founded. Jensen does not argue that within-group heritability entails a non-zero between group heritability. What Jensen *does* argue is that high within-group heritability of IQ (or *g*) among two groups in conjunction with empirical data about the relation of certain environmental variables and IQ (or *g*) makes it more plausible than not that there is a non-zero between-group heritability of IQ (or *g*). In this section I will expose as benign Gould's criticisms to factor analytic methods. I also hope to exonerate Spearman and Jensen from Gould's accusations that they reified *g* in some scientifically illegitimate manner.

Consider the following passage:

> I was taught [factor analysis] as though it had developed from first principles using pure logic. In fact, virtually all its procedures arose as justifications for particular theories of intelligence. Factor analysis, despite its status as pure deductive mathematics, was invented in a social context, and for definite reasons. And, though its mathematical basis is unassailable, its persistent use as a device for learning about the physical structure of intellect has been mired in deep conceptual errors from the start. The principal error, in fact, has involved a major theme of this book: reification—in this case, the notion that such a nebulous, socially defined concept as intelligence might be identified as a "thing" with a locus in the brain and a definite degree of heritability—and that it might be measured by a single number, thus permitting a unilinear ranking of people according to the amount of it they possess (Gould, 1996, pp. 268-269).

In the above passage, Gould contends that factor analytic techniques "arose as justifications for theories of intelligence," but is that so? John Carroll (1995), a leading cognitive abilities researcher and expert on factor analysis, disagrees, "factor-analytic procedures can be regarded as devices to assist in developing different theories of intelligence and choosing among them." Carroll does not regard factor analysis as a justification of his (or any other) theory of cognitive ability.[5]

[5]Thurstone (1938); Guilford (1967); and Cattell (1971).

Specific factorial solutions to correlation matrices need not be interpreted as literally describing cognitive structures. The use of '*g*' to refer to the general factor of intelligence does not thereby commit one to the claim that *g* is a concrete thing with a precise locus in the brain. Factor analysts may, in keeping with a healthful metaphysical prudence, regard *g* (or any other factor for that matter) simply as a "source of variance, dimensions, intervening variables, or "latent traits" that are useful in explaining manifest phenomena, much as abstractions such as gravity, mass, distance, and force are useful in describing physical events," (Carroll, *ibid.*).

However, reification is not always methodological transgression. Sometimes scientists posit entities to "explain" phenomena, in which case the existence those entities, claims regarding their alleged causal powers and properties, are hypotheses that are subject to empirical testing. What Ockham's razor cautions us against is the *unnecessary* postulation of entities. Successful science is rife with theoretical posits, and it is through the course of inquiry that we have come to regard some of these entities as real and others are useful fictions. Clearly, Gould believes that the reification of *g* is an instance of illicit hypostatization. We have already seen that reifying *g* is unnecessary to make sense of the science of cognitive abilities. Gould contends that principal component analysis (PCA), i.e., the statistical method used by Spearman and Burt, is an inadequate basis on which to reify *g*, though Spearman and Burt reify *g* on the basis of PCA nevertheless.[6] But it is unclear than *anyone* in the psychometric community believes that the results of PCA alone are sufficient to justify an entire theory of cognitive ability or warrant the postulation of a theoretical entity. Indeed, it is doubtful that even Spearman reified *g* in spite of Gould's claim that he

> reified [the first principal component] as an "entity" and tried to give it an unambiguous causal interpretation. He called it *g*, or general intelligence, and imagined that he had identified a unitary quality underlying all cognitive mental activity—a quality that could be expressed as a single number and used to rank people on a unilinear scale of intellectual worth (Gould, 1996, p. 281).

[6] Strictly speaking, PCA is not a factor analysis, though its results are known to correlate highly with the results of factor analysis.

It is difficult to square the above quotation from Gould with the following passage from Spearman:

> But notice must be taken that this general factor *g*, like all measurements anywhere, is primarily *not any concrete thing but only a value or magnitude*. Further, that which this magnitude measures has not been defined by declaring what it is like, but only by pointing out where it can be found. It consists in just that constituent—whatever it may be—which is common to all the abilities interconnected by the tetrad equation. This way of indicating what *g* means is just as definite as when one indicates a card by staking on the back of it without looking at its face…Eventually, we may or may not find reason to conclude that *g* measures something that can appropriately be called "intelligence." Such a conclusion, however, would still never be the definition of *g*, but only a "statement about" it (Spearman, 1927, pp. 75-76).[7]

How, in the face of this passage, Gould could surmise that Spearman was guilty of multiplying ontological commitments beyond necessity (or what the evidence warrants) is a mystery. Spearman did suggest that *g* might be thought of as "mental energy," but it seems that in light of the above passage, it might be more appropriate to understand Spearman as metaphorically (i.e., not literally) describing *g* when he talks about mental energy. In any case, the mental energy hypothesis is just that—a hypothesis, subject to test. There is nothing unscientific about hypothesizing. But even if Spearman did reify *g* in some unscientific way, this would hardly be an indictment against contemporary psychometric research. The scientific study of cognitive ability has come some way since the early 20th century. Perhaps if Jensen were guilty of reifying *g*, Gould's complaints may get a toe-hold. After all, in the expanded version of *Mismeasure* Jensen is targeted as *g* reifier. But is Jensen guilty as charged? Consider the following four quotations, the first two from Jensen (1980) (the very work in which, Gould claims, Jensen outs himself as a *g* reifier) and the latter two from Jensen (1998):

> 1. At the strictly empirical or observational level, *g* is best thought of *only as a mathematical transformation* of the correlations among the tests that permits us to

[7] Italics added.

> summarize the raw fact of the test intercorrelations in terms of each test's correlations with *g*, that is, its *g* loading (p. 249).[8]
>
> 2. At present, it seems safe to say, we do not have a true theory of *g* or intelligence, although we do know a good deal about the kinds of tests that are the most *g* loaded and the fact that the complexity of mental operations called for by a test is related to *g* (p. 251).
>
> 3. Although a factor is identifiable and quantifiable, it is not directly observable. It is not a tangible "thing" or an observable event. So we have to be especially careful in talking about factors, lest someone think we believe that we are talking about "things" rather than hypothetical and mathematical constructs…. This does not imply, however, that scientists cannot inquire about the relationship of a clearly defined construct to other phenomena or try to fathom its causal nature (p.55).
>
> 4. It is important to understand that *g* is *not* a mental or cognitive process or one of the operating principles of the mind, such as perception, learning, or memory. Every kind of cognitive performance depends upon the operation of some integrated set of processes in the brain. These can be called cognitive processes, in formation processes, or neural processes. Presumably their operation involves many complex design features of the brain and its neural processes. But these features are not what *g* (or any other psychometric factor) is about. Rather, *g* only reflects some part of the *individual differences* in mental abilities…that undoubtedly depend on the operation of neural processes in the brain. By inference, *g* also reflects individual differences in the speed, or efficiency, or capacity of these operations. But *g* is not these operations themselves (p. 95).[9]

Oddly, Gould does not support his accusations against Jensen with quotations. Rather he simply refers to Jensen's *Bias In Mental Testing* by publication date. The reader is left to take Gould's word on Jensen. However, the relevant chapter of Jensen (1980) does not support Gould's accusations. The chapter specifically on *g* is technically daunting, and the book itself is intended for a professional audience unlike Gould's popular science book, which is intended for a lay audience. It appears that Gould has done his audience a disservice by misrepresenting his opponents.

Certainly the presence of a general psychometric factor is a *prima facie* reason for believing that there is an underlying causal factor operating, responsible for the observed intercorrelations. It is by no means conclusive, but the presence of a psychometric general factor

[8] Italics added.
[9] Italics in the original.

compels us to investigate further to explain the observed correlation. So far I have shown (1) that the use of factor analysis need not carry with it ontological commitments for such commitments are renounced by psychometricians, (2) that Spearman reified *g* even in his early work is not clear, and (3) that psychometricians do not, in general, reify *g*, or any other statistical factor.

Now I will examine the argument that Gould gives for thinking that the presence or derivability of *g* is not even a *prima facie* reason for believing that there is an underlying causal factor operating to produce the perceived intercorrelations. In brief, the argument is that *g qua* theoretical entity *and* statistical factor is underdetermined, therefore *g* is neither real nor does it require an explanation.

Gould presents his case as follows:

> Another, more technical argument clearly demonstrates why principal components cannot be automatically reified as causal entities. If principal components represented the only way to simplify a correlation matrix, then some special status for them might be legitimately sought. But they represent only one method among many for inserting axes into a multidimensional space (p. 282).

In the above passage, Gould is denying the ineluctability of *g* on the grounds that it is not the only way to summarize the information contained in a correlation matrix. There are other statistical methods for achieving the same goal. Gould continues:

> During the 1930s factorists developed methods to treat this dilemma [in finding the correct location of axes] and to recognize clusters of vectors that principal components often obscured. They did this by rotating factor axes from the principal components orientation to new positions…. [But in doing this,] *g* has disappeared. We no longer find a "general factor" of intelligence, nothing that can be reified as a single number expressing overall ability. Yet we have lost no information…How can we argue that *g* has any claim to reified status as an entity if it represents but one of numerous possible ways to position axes within a set of vectors? (p. 283)[10]

This passage raises several important points. The concluding question seems to presuppose that the goal of PCA is to deliver a reified factor. I have already argued that this is not the case. At

[10] Quoted in Carroll (1995).

most, a principal factor *putatively* denotes an entity or cluster of causal processes. Principal factors are abstractions from data, and absent any evidence to the contrary, they ought to be regarded as nothing more than that. I agree with Gould that the evidence provided by a PCA underdetermines the existence of a reified, causal *g* when the results of a PCA are the only evidence. However, I believe that Gould's argument is meant to be much stronger. If what Gould says is correct, then the results of a PCA reporting the presence of *g* are not evidence for the existence of a causally relevant *g* factor, since if we were to rotate the factor axes, *g* disappears. Two equivalently informative methods for describing the data give different results with respect to *g*: that *g* "exists" as the first principal component, and that there is no *g* when the (orthogonal) factor axes are rotated away from the principal component vector. Thus, *g* is simply a statistical artifact. This argument is flawed despite its initial plausibility.

Gould's argument is that the results of a factor analysis cannot be grounds for belief in a theory of cognitive ability whose central theoretical construct is *g* since *g* only arises in the context of certain factor analyses, e.g., hierarchical factor analysis, and not in others, e.g., Thurstone's multiple factor analysis. From this underdetermination, we are to infer that *g* is a mere statistical artifact. But traditionally the lesson to be drawn from underdetermination theses is not that we should *deny* that the entities in question exist, for this saddles one with an empirical hypothesis that is itself underdetermined. Rather, the appropriate attitude toward such entities should be one of cool agnosticism. How can a *g*-theorist such as Jensen respond to this argument?

Jensen (1982; 1998) and Carroll (1995) have in fact responded to this argument. While rotating the principal factors *a la* Thurstone may help to make sense of the factor structure when we have antecedent reasons to think that two different abilities are being measured, the apparent disappearance of *g* is innocuous. Factor rotation maximizes test loadings on one factor while minimizing loading on the other factor—rotating the factors to produce this result while

maintaining the orthogonality of the principal factors is known as the *simple structure criterion*.[11] Factor rotation merely redistributes *g* among the rotated factors. *g* has not actually vanished; evidence of its presence has simply been obscured in favor of elucidating the simple factor structure. Moreover, the fact that scores on mental tests are all positively correlated makes it nearly impossible to rotate orthogonal factors in compliance with the simple structure criterion since some tests will load nearly equally as well on at least two factors, viz., those tests such as Raven's Progressive Matrices that are highly *g* loaded or, less circularly sounding, tests that most closely approximate unidimensionality.

But what if we drop the requirement of orthogonality of factor axes? Thurstone attempted to circumvent the problem posed by the impossibility of adhering to the simple structure criterion with orthogonal factors by inventing oblique factors, i.e., correlated factors. Oblique factors have a separation <90 degrees. In the case of two oblique factors, each is constructed so that it best fits a group of test vectors.[12] However, the correlation between the oblique factors will be a second order factor, namely *g*, when subjected to hierarchical factor analysis. It is ironic that Thurstone eventually came to accept a single factor theory of cognitive ability.

In sum, we have seen that rotation in accordance with Thurstone's simple structure criterion, which is the case that Gould pushes against the ineluctability of *g*, is impossible in the case of mental test vectors since they are all positively correlated. In any case, the rotation of oblique factors simply disperses *g* across the rotated factors. Moving to the case of oblique factors, we see that the application of hierarchical factor analysis yields *g* as a second or third-order factor.

In this section I have shown that Gould's objections to *g* are spurious and miss their mark. First, he presumes that PCA is the only method of extracting *g*, which is false. Were *Mismeasure* published in 1904, Gould's criticism might carry more force, but as it turns out, there

[11] To say that two factors are *orthogonal* is to say that they are uncorrelated.
[12] In the case of two factors I and II, where the angle of separation is X degrees, $r_{I,II}$=cos X.

are multiple methods of factor analysis that Gould ignores and that can be used to extract *g*. Additionally, factor analyses (of any kind) do not commit one to ontological realism regarding those factors. It is hasty to accuse Spearman of naively reifying *g*. Spearman, and Jensen after him, recognized the dangers of reification, and even if the former is guilty of reifying *g*, the latter is not. It is arguable whether Spearman actually reified *g* when he likened it to mental energy, but it is inappropriate to accuse Jensen of reifying *g*. In Spearman's defense, positing an unobservable cause, or "reifying", is scientific commonplace, well within the range of legitimate methodological practices. That one can "rotate away" *g* is not evidence that there is no single general factor that accounts for the variance among scores on a large battery of mental ability tests. In fact, *g* always arises, that is, it is ineluctable.[13] If there are reasons to remain skeptical about *g*, they are not Gould's. In the next section, I will consider Clark Glymour's reservations about factor analysis. There we will find much more substantive concerns.

4. Glymour's Objections

A more serious objection to factor analysis comes from Clark Glymour (1998) in a review of Herrnstein and Murray (1994). Glymour objects to the use of factor analyses in psychometric research on the grounds that it is an unreliable method for the discovery of unobserved causal factors and that it does not deliver univocal factor models of *g*.

Glymour's assessment of factor analysis differs from Gould's. Recall that Gould's principal complaint was that there exist statistically equivalent explanatory structures that should not be taken literally when they are reified to explain observed correlations. Glymour concedes that were his complaint Gould's, factor analysis would find itself in good company, for Gould's objection indicts much of successful science, including Gould's own paleontology, wherein we find different theories positing different theoretical entities which save the phenomena equally

[13] That is, *g* always arises when the battery of tests "is sufficiently large to yield reliable factors and the tests are sufficiently diverse in item types and information content to reflect more than a single narrow ability," (Jensen, 1998, p.73).

well. Glymour's concern is with the kinds of alternatives that factor analytic methods allow, the assumptions employed to eliminate alternatives, and the consequent failure of factor analytic results to represent the correlational data.

Spearman's discovery of *g* founded in his observation that when mental tests i, j, k, and l are all positively correlated, they obey the following vanishing tetrad equation

$$\rho_{ij}\rho_{kl}=\rho_{il}\rho_{jk}=\rho_{ik}\rho_{jl}$$

where ρ_{xy} is the Pearson correlation between performance on tests x and y. Correlations obey the equation when there exists a *latent* common source of variance for each test or in Glymour's words "a single common cause," (1998, p. 5). The residual variance not attributable to the influence of the general common factor is accounted by factors peculiar to each test, i.e., error variance or uniqueness, (hence Spearman's so-called "two-factor" theory). The tetrad equation suggests a factor structure that can be represented graphically for six tests X1, X2, X3, X4, X5, and X6, and their corresponding error terms ε as follows:

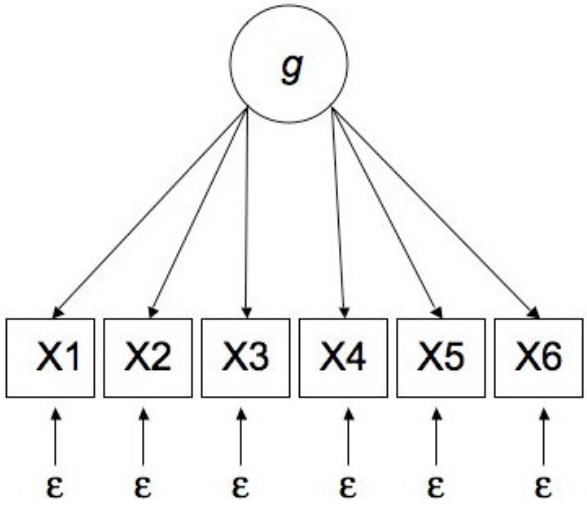

Figure 1.1: Spearman's two-factor theory.

However, the tetrad equation is consistent with alternative factor structures positing more latent causal variables. One such alternative would have a latent cause k as a common source of variance in tests X2−X3, with a second latent cause h accounting for the variance in X1 and k. Still more alternatives exist. Therefore, the factor structure is underdetermined by the data, and factor analysis fails to yield a univocal answer to the question of which factor structure is the correct one.

There is an obvious candidate for eliminating the alternatives that posit latent causal structures beyond the structure that has *g* as the common cause of variance for all the tests, namely a principle of theoretical economy. One motivation for the principle of theoretical economy is the assumption that those theories that posit fewer unobservables are more likely to be true. Glymour does not consider the viability of such a principle for ruling out alternative factor structures, but he does suggest that a much weaker and metaphysically conservative principle will do the job. The principle, which he calls "faithfulness," is the assumption that vanishing tetrads are implied by the underlying causal structure rather than depending on the constraints on the factor loadings of each test. Glymour gives the following two graphs to illustrate this point:

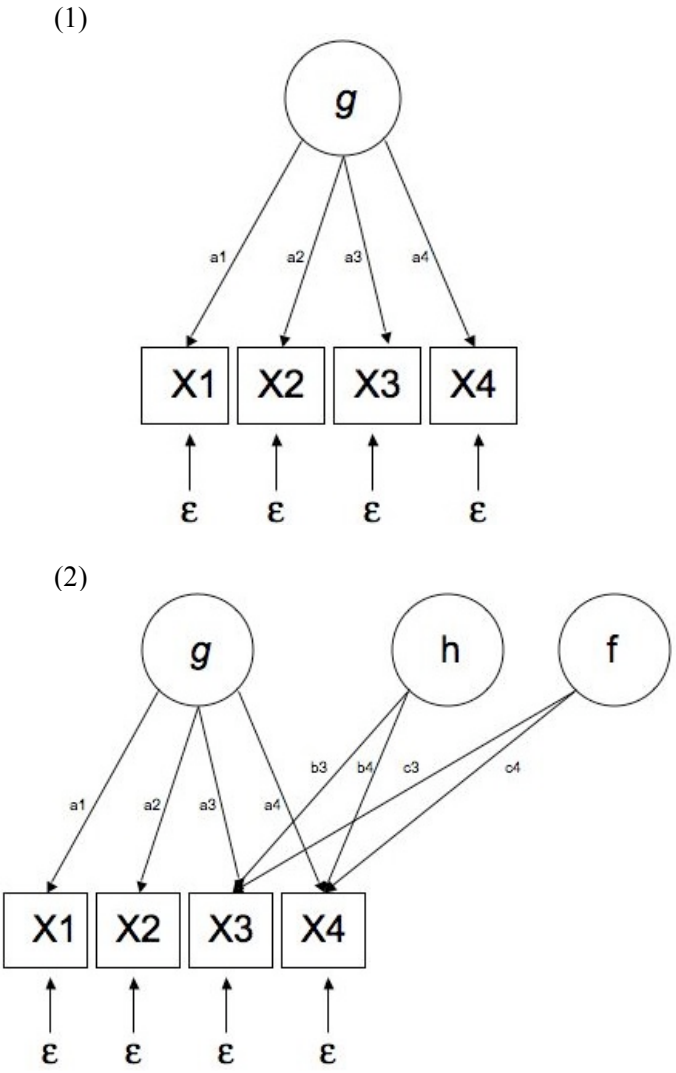

Figure 1.2: Two factor structures.

Let the *g*-loadings in each graph be represented by a_i, and let b_i and c_i represent the factor loadings of h and f in graph 2, respectively, where the indices range over the Xi connected to each factor. In graph one, as in Figure 1.1, the vanishing tetrads are implied by the commutativity of multiplication, i.e. $a_i a_j a_k a_l = a_i a_k a_j a_l$, thus obeying the tetrad equation

$$\rho_{12}\rho_{34}=\rho_{13}\rho_{24}.$$

However, for graph 2 to imply this tetrad equation, the following must hold

$$a_1a_2(a_3a_4+b_3b_4+c_3c_4)= a_ia_ka_ja_l$$

which implies

$$b_3b_4=-c_3c_4.$$

But faithfulness or theoretical economy alone will not save Spearman, for according to Glymour there are alternative structures that imply the tetrad equations that are, in contrast to models such as graph 1 in Figure 1.2, not latent common cause models. For example, consider Figure 1.3:

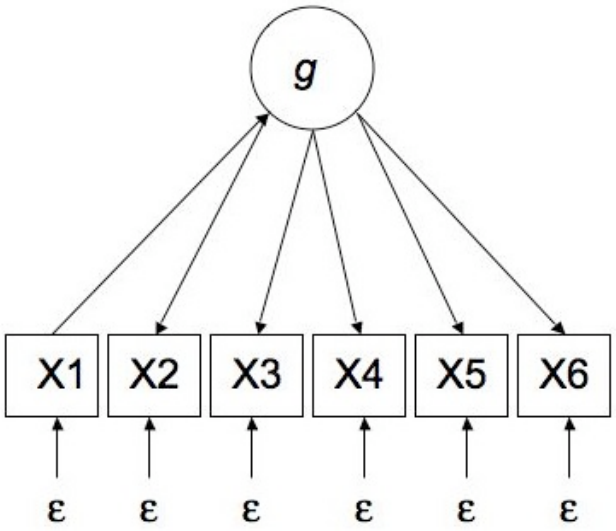

Figure 1.3: Factor model statistically equivalent to Spearman's two-factor theory.

Figure 1.3 is statistically indistinguishable from Figure 1.1, but the common cause is not latent (*g*); the common cause is, rather, the exogenous manifest variable X1, as is indicated by the arrow running from X1 to *g*. Vanishing tetrads are implied by a the presence common cause, though the cause need not be latent, as illustrated above in which X1 is the common cause. In fact, when considering actual tests, the tetrads do not vanish. In order to get the correlations to approximate the tetrad equations in real tests, Spearman's followers introduced additional common causes to account for residual correlations not captured by *g*. These models, which I return to in chapter 4,

are called "bifactor" models. The practice of introducing additional factors to account for the residual variance had the effect of ensuring derivation a *g*-factor solution for a correlation matrix even if other factor solutions, e.g., Figure 1.3, were implied by the data. Glymour contends that Spearman's followers never considered the reliability of this procedure as a method of discovering the causal structure underlying the statistical data.[14]

Thurstone's method for reducing the number of latent common causes was to impose on the correlation a "simple structure," which he (wrongly) took to be unique for any correlation matrix. However, even if there was no unique simple structure for a correlation matrix, we can ask how the criterion fared as a method for reducing underdetermination. Glymour contends that Thurstone's simple structure criterion had no independent evidential support, i.e., there were no grounds for thinking that the underlying causal structures (that is, the mental structures responsible for the data) obeyed Thurstone's simplicity assumption. Thurstone's factor analytic techniques supplanted Spearman's tetrad analysis nevertheless, and for reasons unrelated to reliability.

The Spearman and Thurstone story just given is the prelude to Glymour's main point: There is yet (or as of the date of the publication of Glymour's article) no proof that *any* factor analytic procedure is reliable, not even in an idealized large sample limit. Furthermore, there is no proof that Thurstone's methods will uncover simple structures when they actually exist.

If Glymour's history is correct, then we have good reasons for rejecting Spearman's method of tetrad analysis and for questioning the simplicity assumption of simple structure in Thurstone's methods. We would also find good reason to be skeptical about the validity of *g* as a legitimate scientific construct. Even Jensen (1998) acknowledges this point.[15] Neither tetrad analysis nor Thurstone's method of factor rotation to simple structure are currently preferred

[14] "Reliability was never an issue," (p. 8).
[15] Jensen writes "The disadvantage of Spearman's method is that if his tetrad criterion shows that more than one common factor exists in the tests, his method of factor analysis will not work. If used, it gives an incorrect *g*. The degree of incorrectness depends on the nature of the matrix to which it is applied," (1998, p. 75).

factor analytic methods. The currently preferred model is the higher order factor model (henceforth, the HF model), to which I will return in due course.[16] Tetrad analysis and the method of principal components analysis developed by Spearman's followers are not even, strictly speaking, properly labeled 'factor analysis'. Even if one does not question the reliability of Spearman's and Thurstone's methods, one can still find reason to reject their factorial solutions to correlation matrices. For example, if patterns of intercorrelations among large groups of IQ tests reveal that some tests are more strongly correlated with one another than with others, Spearman's hypothesis is thus falsified, for this would indicate the presence of group factors (in accordance with Thurstone's model). Likewise, if patterns of intercorrelations among large groups of IQ tests exhibit a positive manifold, Thurstone's hypothesis is thus falsified, for this would contradict the hypothesis that there exist uncorrelated groups of primary abilities. These facts suggest the HF model preferred by contemporary *g*-factor theorist: that there is a general factor of intelligence, but that there are also less general group factors. That is, hierarchical factor analysis yields the following confirmatory factor structure:

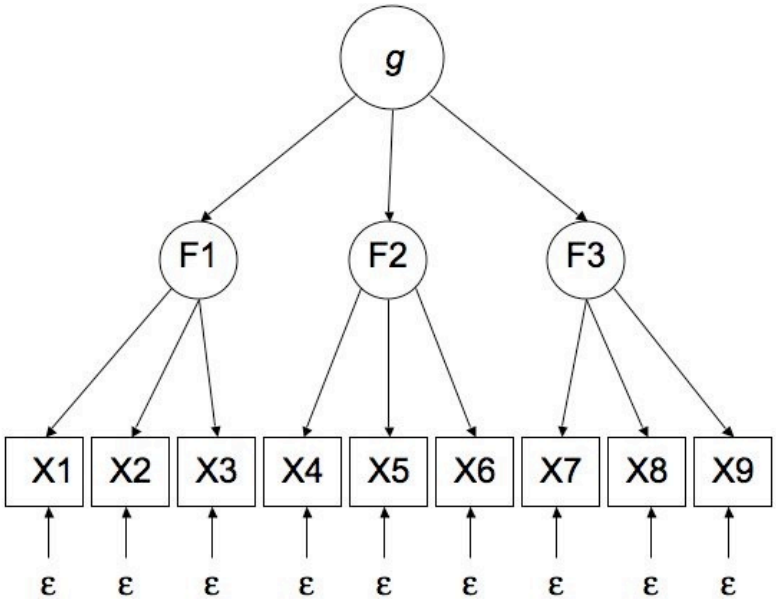

Figure 1.4: The higher order factor model.

[16] This model is also known as the orthogonalized hierarchical model. It is given a detailed treatment in Chapter 4.

This brings us to Glymour's points regarding the reliability of *any* factor analytic procedure and the related issue of underdetermination—that there will be multiple factorial solutions for any correlation matrix exhibiting a positive manifold and that adjudicating between solutions will be grounded in principles external to the (possibly unreliable) statistical procedures that generated them.

To the following extent I agree with Glymour's assessment of the factor analytic approach to the study of the structure of mental ability: factorial solutions themselves provide little evidence for believing specific causal hypotheses regarding the actual structure of mental abilities. The data will restrict the range of possible factorial solutions for a given correlation matrix; however, for any matrix, there will be multiple solutions, and adjudicating between them will require extra-statistical procedures. This much underdetermination is tolerable, for to this extent all scientific theories are underdetermined: physical systems never admit of one model only. The problem arises when there are alternatives implied by the matrix such as in Figure 1.3 in which a measured variable accounts for the latent common cause posited as the source of the variance in the rest of the factors. Therefore, in order to meet Glymour's challenge, it will be incumbent upon me either to discern an acceptable criterion for adjudicating between the alternatives given in Figure 1.1 and Figure 1.3, or to dismiss those factor structures as constituting an objection to the *g*-factor theory. Rather than provide some *a priori* constraint on admissible factorial solutions, I will search for one in actual scientific practice. However, first I will address the reliability concern: are all factor analytic procedures unreliable for discovering latent causal structures?

The question of the reliability of factor analysis for uncovering latent causal structure is germane to legitimacy of the psychometric conception of mental ability only if psychometricians take themselves to be formulating causal hypotheses on the basis of factorial solutions. However, as I showed in the previous section in which I considered Gould's objections, it is not clear that

psychometricians, especially Jensen and Spearman, were doing any such thing. For if *g* is regarded simply as a common source of variance in performance on (groups of) tests of mental ability, then *g* cannot be the sort of thing that can enter into a causal relation. Only if '*g*' denotes some property or cluster of properties can it play the role that Glymour claims psychometricians want it to play. The arrows in Glymour's factorial solutions given above are arrows denoting alleged causal influence, but if *g* is just a statistical factor, then Glymour's directed graphs should not be interpreted causally.

Of course, one may object that if psychometricians do not take themselves to be discovering *via* factorial structures the structure of cognitive ability, what *are* they doing? How do they motivate their inquiry and continued interest? Jensen himself admits that

> factor analysis would have little value in research on the nature of abilities if the discovered factors did not correspond to real elements or processes, that is, unless they had some reality beyond the mathematical manipulations that derived them from a correlation matrix (Jensen & Weng, 1994, p 254).

In nature spurious correlations abound, and some of them enable successful prediction, but they do not further our understanding of natural phenomena. If *g* has a similar spurious status, then the psychometric study of cognitive ability seems not so scientific and, moreover, it seems not so concerned with cognitive ability. In chapter 3 I will address the question of whether psychometricians are committed to some form of realism regarding latent variables and whether they are also committed to the further claim that the referents of latent variables are causally relevant to the values of less general latent variables and, ultimately, observed scores on mental ability tests. I will also address whether the psychometric approach to the study of cognitive ability can be rationally motivated if psychometricians adopt an anti-realist stance toward latent variables. For now, however, I will simply assume that psychometricians *are* aiming to formulate causal hypotheses with factor analytic data—this will give Glymour's objections a target.

Glymour, like Gould, ignores the distinction between two kinds of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Neglecting the distinction between CFA and EFA leaves a lacuna in any exposition of the psychometrician's method and makes them appear hasty and naïve in forming hypotheses about the structure of cognition. EFA is aimed at uncovering factor structures underlying correlation matrices. No prior theory of the factor structure is assumed and the underlying factor structure is inferred from factor loadings on variables. CFA, on the other hand, seeks to confirm a hypothesis regarding the underlying factor structure. By using CFA, researchers test the fit of a previously hypothesized factor structures to data. Glymour writes as though EFA is all there is to obtaining *g*. One may perform EFA initially to explore possible factor structures that underlie the data; however, once EFA has contributed several possible factorial solutions, one can then confirm those factor models using CFA essentially checking goodness of fit between the candidate factor structures and a data set.

One may also object that Glymour has set unreasonably high methodological standards. Correctness proofs, even in the large sample limit, are not necessary to establish a discovery procedure's legitimacy. Lacking a correctness proof itself is not a reason to reject, wholesale, a statistical discovery technique; rather, the appropriate reaction is measured modesty regarding the results of employing such a technique. Exploratory techniques do not yield unique factorial solutions, and given that the solutions may be inconsistent, we should be skeptical regarding resulting factor models *qua* causal hypotheses until the models have been subjected to confirmatory tests. The procedure here outlined also speaks to the objection that there is no acceptable criterion for adjudicating between competing factor models. CFA will further reduce the range of admissible models generated by EFA.

But what about the case of Figure 1.1 versus Figure 1.3? How does the psychometrician adjudicate between the two factor models? Note that this is not a problem unique to psychometrics, for the problem is essentially the classic problem of underdetermination and

confronts nearly every theoretical hypothesis. We have two empirically equivalent theories, i.e., they both generate the same observational (correlational) data. However, that a technique affords one the ability to generate mathematically equivalent alternative theories is no reason to reject the technique lest we reject model theory, logic, and the whole of mathematics. So it is no objection to factor analytic techniques that Figure 1.1 and Figure 1.3 are non-copossible factor models of the same correlation matrix. EFA is a technique for generating theories, CFA is a technique for testing theories generated by EFA, and to the extent that CFA does not deliver a univocal factor model, we should adopt only the reasonable attitude in a case of genuine underdetermination: agnosticism pending further tests. However, Figures 1.1 and 1.3 do not constitute a case of genuine underdetermination since both are falsified by actual data, so we need not worry with them. The mere logical possibility of alternatives is no reason to suspend credence. Fairy tales do not undermine current theory. When Glymour can generate real (in the sense that they are live alternatives for psychometricians and not simply contrivances), non-copossible competing factor models, then he will have given us a genuine case of underdetermination; however, then we should not necessarily reject factor analysis, rather we should adopt an attitude of agnosticism regarding each of the alternatives until further tests warrant adopting a different attitude.

While the fact that EFA does not generate unique solutions does not license the rejection of factor analysis as a legitimate technique, it would be an objection to using the technique if it failed to reveal factor structures when they in fact exist. Glymour makes a similar objection when he writes "[t]here is not even a proof that the procedures find a correct simple structure when one exists," (Glymour, 1998, p. 9). In absence of such a proof, the best defense against this objection is to look at actual scientific studies and see if, despite the lack of proof, *g* is nevertheless robust across different factor analytic techniques. If the presence of *g* is a replicable result and if it is a result that in insensitive to variation in the method by which we measure it, then there seems to be a *prima facie* reason for taking *g* seriously and investigating it further even if there is no *proof* that the results of any particular factor analysis are correct. Note that at this point I am *not*

addressing the question of whether the methods warrant any particular causal hypothesis about cognitive structures. The present concern is conceptually prior to whether a factorial solution warrants a particular hypothesis about cognitive structure. The challenge is whether *g qua* statistic is robust enough to warrant even entertaining the further question of whether *g* is "real" or not. We may lack a proof that there is a *g* when factor analytic procedures tell us that there is one, but if the results of applying different factor analytic procedures to a diverse sample of correlation matrices tell us that there is a *g* when there, in fact, is one and not otherwise, it seems reasonable that novel analyses of correlation matrices that also yield *g* should be taken seriously. What we end up with is not something as strong as the proof that Glymour requires, but, rather, strong inductive support for the claim that factor analytic methods are reliable methods in cognitive ability research, which then motivates further inquiry into the physical interpretation of the results.

Jensen and Weng (1994) take on the challenge of demonstrating the robustness of *g*. They apply six different methods of factor analysis to four contrived correlation matrices whose factor structure is already known. Ten methods of factor analysis are applied to a real correlation matrix derived from the performance of 145 seventh and eighth graders on twenty-four diverse mental ability tests.

Jensen and Weng then compared the actual *g* loadings on the simulated matrix to the loadings reported by the six methods. Some of the simulated matrices were designed to mislead certain methods into reporting *g* loadings that were different from the true loadings. The result was that the correlations between the true *g* loadings and estimated *g* loadings ranged from .997 to .999. Deviations from the true *g* loading ranged from .031 to .059, which is negligible in light the fact that the average factor loading was .50. Thus, *g* appeared robustly and reliably across different methods of analysis and its factor loadings differed little across different methods of analysis. So, not only did *g* always appear, its estimated values were consistently very close the real values.

For the case of the real correlation matrix, the true *g* loadings were unknown, however when the estimates reported by ten different methods of factor analysis were compared and found to have correlations that ranged from .991 to 1.00. Based on the result of the simulated matrix study that strong agreement was present when the tests all closely estimated the true *g* for the correlation matrix, Jensen and Weng inferred that the twenty-four tests, since they strongly agree on what they report to be the factor loadings, also closely estimate the true *g* for the matrix. Just how closely the estimates resemble the true *g* is unknown.

Based on the invariance of *g* across a variety of correlation matrices and extraction methods, and the ability of the methods to accurate estimate the true *g* for a correlation matrix, I infer that despite the absence of a proof that factor analysis will reliably yield a *g* when there is one, factor analysis is reliable enough to warrant further inquiry into the nature of *g*. The choice between different factor models will not be ultimately decided by our statistical methods, and to require that they do may be to demand too much of the mathematical tools. No matrix will admit of a single factorial solution (this is known as the factor score indeterminacy problem), but EFA will yield an array of hypotheses that we can then test using methods of CFA. Searching for neurological correlates of the factors may then further test the resulting subset of empirically adequate models.

In conclusion, Glymour's methodological objections are not above contention and do not undermine research on psychometric *g*. What Glymour's objections *do* show is that factor analytic techniques alone do not give a univocal answer to the question "what is the structure of human cognitive ability?" I have attempted to show that this is a scientifically and philosophically tolerable result that helps to motivate the current project of discerning the theoretical status of latent variables, especially *g*. Indeed, even if Glymour's arguments proved cogent and did pose a serious difficulty for psychometrics, there would still remain the task of ascertaining the philosophical commitments of those psychometricians who take *g* seriously. The fact remains that *g* is one of the most replicable results in psychology (Deary, 2000; Gottfredson, 2002), and the

American Psychological Association task group on intelligence considers the theory of general intelligence to be "the most widely accepted current view," (Neisser *et al*. 1996). For these reasons it seems that interpreting *g* and subjecting to philosophical scrutiny those theories of cognitive ability that might aptly be described as "*g*-centered" is a project worth undertaking. Glymour's criticisms are all at the level of the statistical tools and do not engage the actual psychological research. A more thorough philosophical analysis would investigate both how the statistical tools are used in the research and the psychological interpretation of the statistical results; then, in light of this analysis, one will be in a better position to assess the status of psychometric research on cognitive ability.

By way of these two critiques of psychometrics I hope to have convinced my reader that psychometrics is not dead in the water and that there is much for psychometrics to gain from a little philosophical scrutiny. Gould and Glymour are by no means the only critics of psychometrics, and psychometricians are not blind to the limits of their discipline. There is plenty of scrutiny that comes from within psychometrics, though ironically it has been less influential than Gould's. Joel Michell and Paul Kline, both of whom are psychometricians, are highly critical of their discipline, arguing that there are no scientific measurements in psychometrics. However, this is not a dissertation on fundamental measurement theory and I will be unable to venture into those deep waters. I will briefly address Michell in chapter 3 when I discuss the methodological assumptions in play in psychometrics. I seek to engage psychometrics on its own terms, not "from the outside" so to speak. I hope that the product is a work that will be of interest both to philosophers of science and psychometricians.

5. The Structure of the Dissertation

The plan for the remainder of the dissertation is as follows. Chapter 2 addresses the problem of validity, i.e., psychological measurement. I advocate a test-based approach to validity that is a refinement of Borsboom's concept of validity (Borsboom, van Heerden, & Mellenbergh, 2003,

2004; Borsboom 2005). Chapter 3 addresses methodological commitments in psychometrics. I investigate whether we can make sense of psychometric practice without recourse to realism. Further, I argue that the question of whether scientific realism is justified in the context of psychometrics is settled in the affirmative if there are valid psychometric tests. Chapter 4 is the most psychometric chapter of the dissertation. I compare two competing confirmatory factor models of intelligence data, the bifactor model and the higher order factor model, and I make a case for adopting the latter on both psychometric and theoretical grounds. I argue that only the higher order factor model readily admits of a realist interpretation. I contrast the differential epistemic status of the latent variables in the higher order factor model and suggest two philosophical attitudes that one can take toward second order latent variables. What attitude one adopts hinges on whether one treats the model as a whole as a measurement model or whether one wants to maintain a principled distinction between measurement models and structural equation models. Chapter 5 focuses on the particular challenges posed by psychometric entities, such as the problem of local homogeneity and the relationship between within-subject and between-subject measurement models and attributes. I argue that general intelligence, if it is "real", is not what we may have thought it was, i.e., a property of individuals. It is, rather, a strange sort of theoretical entity that may be related to individual ability, but does not indicate or denote individual ability. This challenge emphasizes the point that absent an understanding of how psychometric attributes function in producing responses on psychometrics tests, realism does little explanatory work. A robust realism not only states that we can be justified in believing that psychometric attributes exist, but it also addresses their causal role in cognitive systems. I also argue that an integrated approach to psychological testing is required to understand the nature of psychological attributes. The resources of psychometrics alone are not sufficient. I argue that if psychometrics is to avoid stagnation, it must actively integrate with collateral disciplines. Avenues for such integration exist and resources for integration are presently available in the fields of cognitive IRT modeling and fMRI studies..

6. A Note on the Broader Impact of the Dissertation

This dissertation aims to deepen our knowledge of the philosophical foundations of psychometrics, a discipline that has been largely neglected by the philosophy of science. In addition to pointing out the epistemic and methodological limitations of psychometric research, I will propose avenues for research and integration with the philosophy of science, the philosophy of mind, and cognitive neuroscience that will enable psychometrics to transcend these limitations. Further, by bringing concepts in the philosophy of science to bear on psychometrics, I will have carved a new niche for the philosophy of science by alerting philosophers to the importance of psychometrics as a scientific discipline worthy of philosophical consideration. Psychometricians will find a new perspective on their discipline and will find reasons to temper their epistemic aspirations with respect to their current measurement models and methods, but will find opportunities for growth as a discipline.

Chapter 2 of the dissertation, which deals with validity, and chapter 4, which concerns the potential mismatch of the structure of interindividual differences and the structure of intraindividual differences, will have far-reaching consequences for how we interpret test scores, particularly scores on tests of mental ability. If the test is valid for measuring general intelligence, a population level phenomenon, it remains to be seen if it is also valid for measuring something in individuals that might be called "general intelligence." General intelligence may turn out to be a construct that cannot be meaningfully ascribed to individuals, even if it can be meaningfully ascribed to or against a background of a population. If this is true, then we will have to rethink how we interpret individual scores on tests that are understood as measuring general intelligence; their validity and utility, while arguably established at the population level, may not carry over to the individual level where we are dealing with dynamical cognitive systems. The account of interindividual attributes (e.g., general intelligence) that I articulate will provide a metaphysical

framework for understanding the kind of thing an interindividual attribute can be, thus offering a constructive revision of how we conceptualize many psychological constructs.

Also worth noting is the potential broader impact of this research outside philosophical and psychometric communities. Intelligence testing and its place in education policy is a topic mired in controversy. Though enthusiasm for intelligence testing has waned over the past two decades, interest in intelligence research seems to be experiencing resurgence as evidenced by recent publications including Zenderland's (1998) *Measuring the Mind*, Bartholomew's (2004) *Measuring Intelligence: Facts and Fallacies*, and John Carson's (2007) *The Measure of Merit.* At a time when there seems to be a market for books on intelligence research, a philosophical analysis of psychometrics and its central claims and methods is especially valuable. It has been eleven years since Stephen Jay Gould's polemical *The Mismeasure of Man* was revised in 1996 in order to answer the then current state of psychometric intelligence and the political fallout surrounding Herrnstein and Murray's (1994) *The Bell Curve*. Many of Gould's objections do not address current psychometric research, and the controversy surrounding *The Bell Curve* has subsided; however, my research will function as a prophylactic against renewed irrational exuberance over the promises of *current* psychometric research as it cautions us against naïve interpretations of the psychometric results.

CHAPTER TWO

VALIDITY IN PSYCHOLOGICAL TESTING AND SCIENTIFIC REALISM

Realists do not fear the results of their study

–Fyodor Dostoevsky

---

1. Validity: Psychology's Measurement Problem

1.1 Introduction

According to some methodologists and psychometricians, validity is the most fundamental concept in psychological measurement. Many of the objections to mental assessment are charges of invalidity—that the tests are biased in some way, that the inferences made from test scores are unwarranted, or that psychological tests do not measure what they purport to measure. The aim of this chapter will be to evaluate two main approaches to validity in psychological testing. Regardless of the conception advocated, validity is central to the question of how to interpret test scores. According to one approach, defended by Denny Borsboom and others, validity is a property of measurement instruments. According to the second general approach, defended by Samuel Messick and others, validity is a property of inferences made on the basis of test scores. Under the first approach there is a subsidiary issue of what features make a test valid. Some would identify these features as metaphysical in character while other would identify them as pragmatic, epistemic, or even moral in character. I will show that whether we take validity to be a property of test scores or interpretive inferences is not important to the issue of whether the belief that psychometric attributes exist can be justified (psychometric realism). I will also show that on one analysis of validity as a property of interpretive inferences, the two conceptions are actually logically equivalent.

If we accept Borsboom's analysis of validity, then issuing a verdict of 'valid' with respect to a test entails commitment to the existence of the attribute purportedly being measured, and thus psychometric realism (realism with respect to psychological attributes) is warranted (at least to the extent the attribution of validity is warranted). Valid tests measure real attributes. On Messick's account, however, a verdict of 'valid' for an inference whose conclusion is 'Test X measures attribute Y' need not entail the *existence* of the attribute purportedly being measured, but it does imply that the belief that the attribute exists is justified, and thus psychometric realism is warranted. This is because validity refers to how justified an inference is and warranted

inferences can lead to false conclusion, i.e., there can be false, justified propositions and beliefs. Despite how it may seem at first glance, this dispute is not mere terminological quibbling. One's choice of terminology does have practical consequences for both the testing industry and clinical psychologists.

The concept of validity in psychometrics has semantic, metaphysical, and epistemological aspects. I will discuss each of these in the following section. Following the analysis of validity, I will relate the concept of validity to scientific realism and argue that the question of scientific realism's tenability in the context of psychometrics is just the question of whether there can be valid psychometric tests. The question of whether any extant psychometric tests are valid will be considered in later chapters.

At issue in the debates about validity are the conceptual foundations of measurement theory in psychology. Though the choice of how best to understand validity has consequences for the testing industry, the legitimacy of the industry is not under dispute. So long as tests continue to provide useful and predictive measurements, the testing industry will retain its market, regardless of fundamental conceptual issues in psychometrics. I will make a case for adopting a conservative concept of validity that does not beg the question in favor of psychometric realism but is nevertheless congenial to it.

1.2 The Semantic Component of Validity

1.21 Borsboom

Denny Borsboom advocates a conservative conception of validity. To motivate his position, it is necessary to understand not only the genealogy of his position, but also to whom he is responding. The classic conception of validity in psychometrics is attributed to Truman Lee Kelley. According to Kelley, "The problem of validity is that of whether a test really measures what it purports to measure," (Kelley, 1927 p. 14). Adopting this conception of validity has ample precedent in psychometric texts and articles (Bartholomew, 2004; Borsboom, 2005; Borsboom *et*

al., 2004, 2003; Cattell, 1946; Cronbach, 1984; Gregory, 1999, 2004; Kelley, 1927; Kline, 1976, 1993; Mackintosh, 1998; Sattler, 2001; Shepard, 1997). For example Gregory (2004, p. 97) asserts, "the validity of a test is the extent to which it measures what it claims to measure," and Kline (1993, p. 34) states, echoing Cronbach (1949, p. 48), that "A test is said to be valid if it measures what it purports to measure."[17] Interestingly, Kline, a prominent psychometrician, claims that his and Cronbach's conception of validity is the "standard textbook definition" and that "the only modification of this definition that [he] is aware of is that of Vernon (1963), who pointed out that a test is valid for some purpose," (Kline, p. 34). There is a variant of the classical conception, however, which emphasizes *inferences* from test scores as being the salient feature of valid tests rather than whether the test successfully measures the target attribute. In *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985, 1999) validity, again, is said to be a property of tests: "A test is valid to the extent that inferences made from it are appropriate, meaningful, and useful."[18] This inferential variant is very close to the other concept of validity that I will discuss.

Some have suggested that the classic conception is something of a category mistake, for the proper bearers of the predicate 'is valid' are *interpretative inferences* made from test scores, i.e., not the tests themselves (Cronbach & Meehl, 1955; Markus, 1998; Messick, 1989a, 1989b, 1995, 1998).[19] A test score interpretation is a claim about the significance of a test score or set of test scores. Markus (1998) considers the concept of validity as applied to tests to be "antiquated," (p. 17). Messick (1989, p.13) claims:

---

[17] In Cronbach's (1949) *The Essentials of Psychological Testing*, the author claims that validity is a property of tests (as quoted), which he maintains in the 1984 edition of the book; however, in the meantime, he published his famous article with Meehl (Cronbach & Meehl, 1955) where they maintain that validity is a property of interpretations of test results, not the test themselves. It is unclear whether Cronbach vacillated between the two notions or if what we get in the 1984 edition is just held over from the 1949 edition.

[18] Quoted in Gregory (2004, p. 97). This characterization of validity is hopelessly problematic since the terms 'meaningful', 'useful' and 'appropriate' are vague, but the feature to which I wish to call attention is the attribution of validity *to tests*.

[19] All subsequent citations of Messick from 1989 will refer to Messick (1989a).

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores and other modes of assessment…. Broadly speaking, then, validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use. Hence, what is to be validated *is not the test or observation device as such but the inferences derived from test scores or other indicators*—inferences about score meaning or interpretation and about the implications for action that the interpretation entails.[20]

And later on the same page, he writes,

> To validate an *interpretive inference* is to ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well-supported…validity is a unitary concept. Validity always refers to the degree to which empirical evidence and theoretical rationales support the *adequacy and appropriateness of interpretations* and actions based on test scores.[21]

In another publication, Messick writes that "[w]hat needs to be valid are the inferences made about score meaning, namely, the score interpretation and its action implications for the test used," (1998, p. 37); and Cronbach and Meehl cavalierly assert in their (1955) that "[i]n one sense, it is naïve to inquire 'Is this test valid?' One does not validate a test, but only a principle for making inferences," (p. 297).[22] These theorists, despite differences in the details, agree that interpretations, not tests, are valid and validity comes in degrees.

In what follows, I will rehearse the arguments of Borsboom *et al*. and their analysis of the meaning of 'validity'. I will then critically evaluate these arguments and consider the semantics of validity on Messick's account. In the subsequent chapter on psychometric realism, I will subject Borsboom's analysis to refinement to arrive at an acceptable formulation of validity.

Borsboom *et al.* (2003) advocate a test-centered analysis of validity (henceforth 'TA-1') according to which a valid test measures what it purports to measure:

---

[20] Italics added.
[21] Italics added.
[22] Cronbach and Meehl advocate an approach to validity distinct from Messick's, but similar in that they agree that validity is not properly said to be a feature of measurement instruments. For Cronbach and Meehl, validity concerns interpretations of test scores and whether they fit within a theoretical nomological network. Psychological constructs are defined implicitly by their place within the network.

(TA-1) Test X is valid for the measurement of attribute Y if and only if the proposition 'Scores on test X measure attribute Y' is true (p. 3).

Validity on TA-1 is a binary property of tests—either tests are valid or they are invalid for the measurement of an attribute Y. TA-1 can be contrasted with what Borsboom claims is the more popular Interpretation Analysis (henceforth 'IA'), which he associates with Messick (1989).[23]

(IA) The test score interpretation 'Scores on test X measure attribute Y' is valid if and only if the proposition 'Scores on test X measure attribute Y' is true (Borsboom et al., 2003 p. 3).[24]

If we apply the above analyses to an IQ-test, say the Wechsler Adult Intelligence Scale (WAIS), and the attribute *general intelligence* we get

(TA-1-WAIS) The WAIS is valid for the measurement of general intelligence if and only if the proposition 'Scores on the WAIS measure general intelligence' is true,

and

(IA-WAIS) The test score interpretation 'Scores on the WAIS measure general intelligence' is valid if and only if the proposition 'Scores on the WAIS measure general intelligence' is true.

According to Boorsboom *et al.*, one advantage of TA-1 over IA is that the latter makes the concept of validity redundant, but the former does not. Validity amounts to truth of the score-interpretation according to IA. Moreover, TA-1 agrees with the classic conception of validity (Kelley, 1927) and the thinking of many contemporary psychologists.

Now, one's choice of analysis may just seem to be a matter of terminological preference, since according to Borsboom it is not because IA suffers from conceptual difficulties that should

[23] It is unclear why Borsboom claims that Messick's analysis is more popular view than TA-1 given the abundance of mainstream and influential psychometric texts, enumerated at the beginning of this section, that take validity to be a property of tests, not the degree of support for interpretations of test scores.
[24] This formulation is actually a generalization of an instance given in Borsboom's article, namely "The test score interpretation 'IQ-scores measure intelligence' is valid, if and only if the proposition 'IQ-scores measure intelligence' is true. However, it is not a generalization to which Borsboom would object; see the following quotation.

compel us to adopt TA-1; rather it is for considerations of terminological parsimony and institutional tradition that TA-1 is preferable to IA. This is not to say that TA-1 is unproblematic, but if there are problems with this account, they are not at the level of the semantics of 'validity'.

There are several reasons to think that TA-1 and IA are not genuine alternatives. As formulated, the test-based approach and interpretation-based approach are logically equivalent. Let T='Test X is valid for the measurement of attribute Y', let P='The test score interpretation 'Scores on test X measure Y' is valid', and let S='Scores on test X measure attribute Y' is true'. We can symbolize TA-1 and IA in the following manner:

(TA-1) $T \leftrightarrow S$

(IA) $P \leftrightarrow S$

But this implies

$T \leftrightarrow P$,

and thus,

TA-1 $\leftrightarrow$ IA.

So Borsboom's formulation of the problem seems to confirm what may have been one's initial suspicion regarding this dispute: TA-1 and IA are but two ways of saying the same thing and that nothing conceptually significant rides on whether validity is treated as a property of measurement instruments as opposed to a property of score interpretations. Since IA and TA are logically equivalent, IA cannot be TA-1's rival. The choice between IA and TA reduces to a terminological quibble.

Borsboom's intended target in objecting to IA is Messick's analysis of validity; however, it is doubtful that Messick would accept either IA as formulated by Borsboom or the claim that validity of an interpretation amounts to truth, even for the particular interpretation given in the formulation of IA. IA certainly cannot be read off of the quotations from Messick. Borsboom does not justify his formulation of IA except by saying,

> But what does it mean to say that a test score interpretation is valid, if not that the proposition that expresses this interpretation is true? That is, there seems little harm in a restatement of validity as [IA] (Borsboom *et al.*, 2003 pp. 2-3).

This is not much of a justification, but it is all Borsboom gives us. The concept of validity is epistemic in character for Messick; it is not a semantic or metaphysical concept. It refers to the degree of empirical support that an interpretive inference enjoys. Truth, on the other hand, is not an epistemological concept for scientific realists such as Borsboom. Furthermore, there seems to be no textual support for attributing to Messick an epistemic theory of truth. Therefore, Borsboom's assumption 'according to IA, to say that an interpretation is valid is tantamount to saying that it is true' is illegitimate. To see how the two concepts, i.e., truth and validity, can come apart one need only consider the possibility of an unwarranted inference to a true conclusion. An interpretation may be true, but it may have scant evidence in its favor, thus we would have a case of an invalid, albeit true, interpretation. In the other direction, false claims might enjoy great inductive support. Ampliative inference is not truth preserving, and it is the nature of interpretations that they are the result of ampliative inferences.

Second, IA, as stated, seems to conflict with the idea that validity comes in degrees, which leads me, once again, to question whether IA fairly represents Messick's position. Only if Borsboom takes Messick to be conflating epistemology and semantics while adhering to a notoriously problematic and marginal theory of truth according to which truth comes in degrees can I understand why Borsboom might have formulated IA as he did. Let us allow that validity comes in degrees under IA for the sake of argument. Borsboom certainly was not ignorant of this aspect of Messick's theory of validity.[25] Interpretations that are not "valid to the maximum degree" are not, strictly speaking, valid, just as a gas tank that is not filled to capacity is not full (and is only such-and-such percent full). Perhaps Borsboom is merely describing what 'validity' means in the limit for the interpretation analyst. If validity refers to the degree of empirical

[25] Personal correspondence with Borsboom supports this claim.

support for an interpretation, then a maximally empirically supported interpretation just is a true interpretation, and all other interpretations are true only to the degree to which they are valid. But this rationale attributes to Messick a theory of truth that he would not accept.

Messick does not conflate semantics and epistemology in a way that would suggest that he would be willing to accept IA. But if IA is a position without an advocate, then it is not a genuine alternative to TA-1. Even worse, aside from its rhetorical function of motivating Borsboom's own view, a conflation of semantics and epistemology is IA's only motivation. So let us now turn to Messick's actual view.

1.22 Messick

That Messick does not take validity to be a property of measurement instruments *à la* Borsboom is obvious from his quotations above. On Messick's account, validity is the degree of confirmation for an interpretation of test scores or a test score. Thus, the epistemic and semantic components of Messick's view are inextricably commingled. There is a further complication. Recall that for Borsboom, there is but one interpretation relevant to assessments of validity, viz., whether a test measures the attribute it is purported to measure. Messick's account of validity is not so restrictive. *All* interpretations are implicated. For any interpretation, not just those that concern the measurements of specific attributes, the justificatory grounds may be assessed with respect to their validity. Some interpretations will concern utility; others may concern ethical consequences of implementing a test. For my purposes, however, I will focus on interpretations relevant to what is traditionally referred to as "construct validity". I will then consider the more general issue of whether a restrictive conception of validity, such as Borsboom's, is preferable to Messick's relatively inclusive account.

One unfortunate feature of Messick's (and Cronbach's and Meehl's) conception of validity, at least for the philosopher, is that while it attributes validity to *inferences* made on the

basis of test scores, Messick's conception of validity is not the logician's.[26] This is a potential source of confusion: to have two logically distinct properties of inferences that go by one name. An interpretive inference can have a high degree of validity (in Messick's sense) without being valid in the sense of *deductively valid*. For example, since the evidence marshaled in support of an interpretation is going to support the interpretation inductively in most if not all cases, such inferences will rarely, if ever, be deductively valid. Now consider the converse. Not all deductively valid arguments whose conclusion is an interpretation will have a high degree of validity (in Messick's sense). For example, if the set of premises is inconsistent, or if it contains as its sole member the interpretation that is also the conclusion of the inference, we would be disinclined to say that the interpretive inference has any validity (in Messick's sense). This is not a damning feature of Messick's analysis, but it is a potential source of confusion.

A more serious cause for concern is a *prima facie* inconsistency in Messick's account. Up to now, I have proceeded with the idea that validity is a *property* of inferences drawn from tests scores and other lines of evidence. This is how Messick's analysis is usually interpreted. There is textual support for this in Messick (1989, p. 13):

> Validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores.

But this is not the only account of validity that we get on page 13:

> Validity is an integrated evaluative *judgment*[27] of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment.

According to this characterization, validity is a judgment, i.e., *not* a property. On the same page, there is yet another characterization:

[26] Cronbach and Meehl (1955) write on p. 297, "If a test yields many types of inferences, some of them can be valid and others invalid." For them, 'valid' means warranted or justified, not deductively valid.
[27] Italics added.

> …validity is an *inductive summary* of both the existing evidence for and the potential consequences of score interpretation and use.[28]

Therefore we have on one page alone three mutually inconsistent accounts of what 'validity' means, and nowhere do we get an account of what 'valid' means. Thus Messick leaves the semantics of validity ultimately unclear. However, the account of validity as a judgment can be dismissed on grounds of incoherence. Validity, for Messick, is essentially degreed. Judgments, on the other hand, cannot meaningfully be said to come in degrees. Moreover, validity is a property and a judgment is an evaluation. Evaluations are not properties; hence, the concept of validity *qua* judgment is mistaken. The same can be said of the third account: summaries do not come in degrees, nor are they properties; therefore, validity cannot be a summary. Summaries can be more or less informative, just as paraphrases can be more or less complete. Informativity and completeness may come in degrees, but the very objects exemplifying these properties do not. Thus a charitable interpretation of Messick's position ascribes to him the first of the three characterizations of validity.

The inconsistency notwithstanding, notice that, at the level of semantics, scientific realism is presupposed neither on Messick's account nor Borsboom's account. Either concept of validity is compatible both with realism and antirealism. A commitment to the reality of psychological attributes such as general intelligence does not follow from either account. The antirealist, if he accepts Borsboom's analysis, will say of every test that it is invalid (or that attributions of validity are unjustified). If he accepts Messick's analysis, will say of every interpretive inference that it is invalid if its conclusion is that a test measures some real psychological attribute. Nevertheless, it is unlikely that antirealists would find either analysis appealing, especially since validation research is rendered unintelligible if one is antirealist since it is aimed at justifying attributions of validity. While neither analysis decides the question of

[28] Italics added.

realism, it is obvious that they are congenial to realism and probably are motivated by realist sympathies, especially Borsboom's. Nevertheless, both philosophical positions are compatible with either concept of validity. The virtues of Borsboom's analysis vis-à-vis Messick's analysis will become clearer once the metaphysics, epistemology, and pragmatics of validity have been considered.

1.23 Pragmatic Concerns

The choice of terminology may *seem* inconsequential to test-developers, theoreticians and clinicians. For test-developers the choice of terminology is especially relevant since adopting TA-1 saddles them with the burden of constructing valid tests, whereas Messick's analysis makes validity a property of inferences made from test scores; thus, it encumbers those who interpret test scores with burden of establishing validity. The theoretician escapes unburdened at the level of semantics. A restrictive account, such as Borsboom's, may issue verdicts of 'invalid' where a more permissive account of validity, such as Messick's, might issue verdicts of 'valid'. One may even make inferences that are valid (in Messick's sense) from scores on tests that are invalid (in Borsboom's sense). On Messick's account invalidity is the problem of those who interpret the scores: educational psychologists, educational institutions, clinicians, even those who construct the tests in as much as the test is designed to license certain inferences about the examinees. On Messick's account, invalidity is not a defect in the measurement instrument (since inferences, not tests, are validated); it is a defect in reasoning from test scores. It might be that an inference suffers from validity problems because of poor test design, in which case we have discovered why the reasoning suffers from validity problems, but it is the inference based on test scores and not the test itself that suffers from low validity.

1.3 The Metaphysical Component of Validity

1.31 Borsboom

In a later publication, Borsboom *et al.* (2004) offer another analysis of what it means to say that a test is valid:

> (TA-2) A test is valid for measuring an attribute if and only if (a) the attribute exists, and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure (p. 1061).

TA-2, like TA-1, is based on Kelley's (1927) analysis of validity; however, it is different kind of analysis than the one Kelley offers or TA-1. While TA-1 is a semantic analysis of validity (though not devoid of metaphysical import), TA-2 is a metaphysical analysis in that it states what empirical conditions must obtain for a test to be valid.[29] That is, TA-2 gives the truth conditions for the *analysans* of TA-1. ''Scores on test X measure attribute Y' is true' is true just when Y exists and variations in Y produce variations in scores on X; therefore, the semantic analysis may be, to some extent, redundant given the metaphysical analysis. However, the metaphysical analysis goes beyond the semantic analysis. TA-2, unlike TA-1, couches validity in causal terms. TA-2 is a metaphysical interpretation of TA-1. Since there could be more than one possible interpretation of TA-1, i.e., some that are couched in causal terms, some that are not, it is unlikely that the semantic analysis will be able to be made fully redundant; the semantic analysis settles the metaphysics of validity in part only.

1.32 Messick

It will come as no surprise that Messick's account of validity is relatively silent on matters metaphysical. Validity is, for him, an epistemic concept after all. To say that some interpretive inference is more valid than another does not commit one to the existence of psychological attributes, nor does his account commit him to any particular theory of psychological

[29] Specifically, TA-2 states what it is to *measure* an attribute.

measurement. The closest Messick comes to doing metaphysics overtly is in his discussion of the interpretation of test behaviors in terms of constructs. He offers a compromise between constructivist interpretations and realist interpretations which he calls "constructive realism". But constructive realism is itself metaphysically evasive. The constructivist view entails that psychological attributes have no existence independent of our efforts to measure them; specific attributes are convenient fictions, mere classifications of behavior. The realist view is metaphysically profligate, for it entails that psychological attributes that tests purport to measure actually exist. Constructive realism is offered as a middle ground: it acknowledges that constructs are mental constructions that we use to make sense of behavior, but those constructs can bear a reference relation to real attributes (or "traits" as Messick calls them):

> This perspective [i.e., constructive realism] is realist because it assumes that the traits and other causal entities exist outside the theorist's mind; it is constructive-realist because it assumes that these entities cannot be comprehended directly but must be viewed through constructions of that mind. By attributing reality to causal entities but simultaneously requiring a theoretical construction of observed relationships, this approach aspires to attain the explanatory richness of the realist position while limiting metaphysical excesses through rational analysis. At the same time, constructive-realists hope to retain the predictive and summarizing advantages of the constructivist view (p. 29).

But it is not clear that Messick's position is consistent with the way constructs are often treated in applied psychology where 'attributes' and 'constructs' are used interchangeably. The quote above admits that constructive realism presupposes that attributes exist, but later, on the same page, Messick writes:

> Nonetheless, this treatment of the constructive-realist viewpoint is not meant to imply that for every construct there is a counterpart reality or cause in the person or in the situation of interaction.

The practicing psychologist is pressed to attempt to resolve the inconsistency by carving apart constructs and attributes. One may agree with the claim that not every construct necessarily has a referent while interpreting the realist component of constructive realism as saying that when they

do successfully refer, they refer to causally efficacious psychological attributes. This resolves the inconsistency witnessed in practice, but it doesn't come for free. Messick's account seems committed to semantic realism, the thesis that scientific claims should be read literally. What is lost is the constructivist thesis that constructs are only convenient fictions, taxonomic tools for the classification of behavior. It is a realist thesis that when constructs enjoy referential success, they denote real attributes. But semantic realism is not the divisive issue among contemporary realists and antirealists. What sets antirealists such as Bas van Fraassen or Kyle Sanford (who are semantic realists) apart from realists such as Richard Boyd or J. D. Trout is the tenability of epistemic realism. Realists, but not antirealists, typically assert that attributions of referential success can be warranted by evidence.

The connection between constructive realism and validity is not obvious from Messick's discussion. The connection is this: by deciding what sorts of entities one will allow into one's ontology, the possibility of accruing evidence for ontological claims is affected. For example, constructivists will say of any interpretive inference that claims a construct refers to a causally efficacious psychological attribute that it has a very low degree of validity because constructs do not refer such things on that account. Whether constructive realists issue a similar verdict will be made not on the basis of semantics, but on the basis of other considerations such as the tenability of epistemic realism. In Messick's, though not in Borsboom's, account the concept of validity carries with it no ontological commitments.

1.4 The Epistemic Component of Validity

In this section I will discuss evidentiary aspects of validity. First I will address the concept of reliability in psychological testing since there seems to be some confusion in the literature concerning its relation to validity. Some authors claim that there is a deep conceptual connection between the two, e.g., that a test's reliability is a necessary condition for its validity. I will argue that this is not the case if we accept Borsboom's conception of validity and that this necessity

claim conflates the metaphysics of validity with the epistemology of validity, but first I will give a short primer on reliability.

Reliability is consistency of measurement (Gregory, 2004 p. 78). Theoretically, reliability is defined as the squared correlation of test scores and true scores in a population of subjects (Lord and Novick, 1968, sect. 3.4). The equation typically given for reliability is

$$Rel(X)=Var(T)/Var(X)$$

where 'Rel' is reliability, 'Var' is variance and 'X' and 'T' respectively denote observed scores and true scores in a population (Mellenbergh, 1996, p. 294).

Because we never have access to "true scores", psychometricians have developed numerous ways to estimate reliability, i.e., operationalizations of the concept of reliability. At a general level, reliability estimates come in two varieties: diachronic and synchronic.

Diachronic: test-retest, alternate forms (delayed).

Synchronic: alternate forms (immediate), split-half, coefficient alpha (also called "Cronbach's alpha").

Diachronic reliability refers to temporal stability of scores for a sample of test-takers. Synchronic reliability, with the exception of the alternate forms (immediate) to be discussed below, refers to the internal consistency of tests.

Diachronic Reliability

Those measures of reliability that are diachronic involve two administrations of tests separated by some interval of time. Test-retest reliability involves the administration of the same test twice to a

population of test-takers. The correlation of performance on the test administered first with the (same) test administered second is the reliability coefficient. Alternate forms (delayed) reliability also refers to the correlation between performance on two tests administered at different times to the same population; however, the second test is different from the first. The second test may have different questions, but the two tests are meant to be similar in both content and difficulty.

Synchronic Reliability

Synchronic measures of reliability involve one administration of a test. Alternate forms (immediate) reliability is demonstrated by administering two forms of the same test to a sample of test-takers some in the same testing session. As before, both forms of the test are similar in content and difficulty. Performance on the two forms is then correlated to yield a reliability coefficient. Because two different tests are given, alternate forms (immediate) cannot be thought of as an index of internal consistency. Internal consistency can be meaningfully applied to a single test form only. Split-half reliability, an index of internal consistency of a test, is demonstrated by administering at one time one form of the same test to a sample of test-takers. The test has two halves that function much like alternate forms of a test: they are, ideally, similar in content and difficulty. Performance on each half is then correlated to yield a reliability coefficient. Coefficient alpha is a generalization of split-half reliability and is, therefore, also an index of internal consistency of a test. Coefficient alpha is the mean of all possible split-half reliability coefficients after correction (Gregory, 2004, p. 86).[30] Coefficient alpha is given by the following formula:

[30] Because each half of a test is typically half as long as an entire test form and shorter tests tend to be less reliable than longer tests which better sample the universe of possible items, the reliability of the entire test will be artificially attenuated by the relatively diminished reliability of each half. To correct for this underestimation of reliability, the Spearman-Brown formula is applied. The formula is

$$r_{SB}=(1+r_{hh})/2r_{hh}$$

where r_{SB} is the corrected reliability, r_{hh} is the reliability coefficient computed by correlating each half of the test (Gregory, 2004 p. 85).

$$r_\alpha=(N/N-1)(1-(\Sigma\sigma_i^2/\sigma^2))$$

where r_α is the coefficient alpha, N is the number of test-items, $\Sigma\sigma_i^2$ is the sum of the variances of the individual test-items, and σ^2 is the variance of test scores.

Interscorer reliability, a measure of consistency of scoring across examiners, applies only to those tests whose scoring requires the examiner's judgment. It is measured by correlating the scores of tests to assess objectivity of scoring practices. A low interscorer reliability estimate would not indicate a shortcoming of the physical test form it would indicate a shortcoming in scoring practices, for features of the examiners, not the test form, are responsible for the low reliability estimate. For these reasons one might object that interscorer reliability is a red herring among the other reliability indices. However, interscorer reliability does belong among the others since the scorer's judgment is an essential part of some measurement instruments. The scorer's judgment is as much a part of the test as are the individual test items.

Each method for assessing reliability has weaknesses and for this reason rarely do test-developers rely on one method alone. Coefficient alpha, because it is easy to calculate and is in an inexpensive way of estimating reliability, remains the most popular.[31] Because it confounds error and change in test performance due to psychological change between testing occasions, assessments of test-retest reliability are usually supplemented with estimates obtained using one or more of the other methods mentioned above.

1.41 Borsboom

It is often claimed that reliability of a test is a necessary condition for the validity of that test (Gregory, 2004 p. 97), though some have denied this (*cf.* Kline, 1976 p. 49; 1993 p. 27). If we

[31] I thank Gideon Mellenbergh and Denny Borsboom for pointing this out to me.

accept TA-2 according to which a test is valid for measuring an attribute A iff the attribute exists and variations in A causally produce variations in measurement outcomes, then the claim that reliability is necessary for validity comes out sounding like a mistake based on a conflation of metaphysics and epistemology. There is nothing in Borsboom's account that suggests consistency of measurement is necessary for its validity. That is, being a reliable test is not a precondition for it validity; it is, at most on his account, a precondition for being justified in believing that a test is valid. An unreliable test may nevertheless measure a psychological attribute. If reliability were necessary for validity, then the sentence 'Test X measures attribute Y unreliably' would imply a contradiction. However, if we are to know that a test is valid, then reliability seems to be required. Radically unreliable tests would be epistemically indeterminate with respect to what they measure. Consider an analogy. Suppose you have a scale that registers a reading for weight only when someone steps on it. However, the scale gives radically different weights for the same person (whose weight does not change) over repeated trials. Measurement outcomes are the result of variations in the attribute weight, but since the readings are not consistent, you cannot know if it is weight that is producing the readings or if it is something else such as the number of beliefs one is currently entertaining, one's lucky number for that moment, or the number of air molecules in one's lungs. Reliability does not seem necessary for the scale to be a valid test of weight, but *correctly interpreting* the scores seems hopeless without it, and if we cannot make sense of the scale's readings, then we seem to have an epistemological obstacle to being justified in saying that the scale is a valid test of weight.

In the context of Borsboom's account of validity, high reliability's place is as merely a normative epistemological constraint on attributions of validity to tests designed to measure stable attributes. It is difficult to see how evidence could accrue in favor of the validity of an unreliable test. Were a test very unreliable and were there no other evidence that the test is measuring what it purports to measure, we could not know if it is measuring its target attribute. Substantive theoretical knowledge of the attribute is needed in order to know how to interpret the

reliability data. Reliability coefficients are correlations, and it is difficult to see how correlation on their own could tell us anything about what a test measures. At best, high reliability is only circumstantial evidence that the same attribute is being measured on different testing occasions. I say "at best" because this interpretation of the reliability coefficient is justified only against the background the substantive theoretical claim that the attribute is stable over time. Without that piece of theoretical knowledge, the reliability coefficient is not interpretable. Reliability seems irrelevant to validity. Reliability coefficients have a home in validation studies, but validation studies merely provide evidence for the validity of a test; neither validation nor reliability is constitutive of validity. Validation procedures are conceptually distinct from validity on Borsboom's account, and he admits as much:

> Therefore, I would like to push my validity conception one step further, and to suggest not only that epistemological issues are irrelevant to validity, but that their importance may well be overrated in validation research too (Borsboom 2005, p. 164).

For Borsboom, validation research gets the process of test analysis backwards. We don't give tests and then determine if we are measuring what we intend to measure. Rather we ought to construct tests with knowledge of the processes we intend to measure and then construct instruments to measure them. If we know the causal facts relevant to an attribute in advance, we will probably have a good idea how to measure it. The challenge is in knowing the attribute we intend to measure.

There are reasons to have reservations about TA-2. Borsboom's concept of validity *is* attractive: it is straightforward, simple, and tidy. But the devil is in the details, and it seems that Borsboom is avoiding any dealings with the devil. He gives us an understandable set of necessary and sufficient conditions for validity, but he is relatively silent on evidential matters. For, example, he does not indicate the epistemic standards of a warranted attribution of 'valid' to a test, nor does he specify the grounds for formulating causal hypotheses. Causal hypotheses are supposed to ground test development, but absent performance on psychometric tests, what will

inform the formulation of causal hypotheses? However, this might be expecting too much from Borsboom. For if he did attempt an account of when an attribution of validity is warranted, he would have, in effect, be arguing that some form of scientific realism is true. If an attribution of validity *à la* TA-2 is warranted, then so is the claim that certain theoretical entities exist, *viz.*, psychological attributes. I prefer to see this omission in Borsboom's theory as an opportunity for further research rather than a shortcoming in his account. After all, it is clear from reading his work on validity that he is more concerned with the conceptual and metaphysical aspects of validity and less so with the epistemological and methodological aspects.

Just as Messick's view is short on metaphysics, Borsboom's is lean on epistemology. This is a consequence of the characters of their analyses. Messick's concept of validity is epistemic; Borsboom's is metaphysical.

1.42 Messick

As we have seen, Messick's account of validity is epistemic in character. Validity refers to the degree to which an interpretive inference is warranted by evidence. Given his relative silence on what it *means* for an inference to be valid and given that his account of validity focuses on how evidence accrues in favor of such inferences, Messick's account is actually better conceived as an account of *validation* rather than validity. It is telling that in the 4th edition of *Educational Measurement*, there is not a chapter entitled 'Validity'. Instead, there is a chapter entitled 'Validation' by Michael Kane (2006). Messick spends a considerable amount of time discussing the sorts of data that count as evidence for an interpretive inference, but never does he specify what conditions must be met in order to justifiably assert of an interpretive inference that it is valid (Kane (2006) is similar in this respect). This is an important point of contrast with Borsboom who makes explicit what it takes for a test to be valid, and he also specifies some minimal epistemic requirements for justifiably asserting that a test is valid. The two also differ with respect to their attitudes regarding validation. For Messick, validity refers to the outcome of

test validation, and so test validation occupies a role of central importance in his theory of validity. Borsboom, as we have seen, believes that the role of validation has been overemphasized. Messick believes reliability is a requirement for validity whereas Borsboom formulates a concept of validity that does not require high reliability, though I've shown that *attributions* of validity do require reliability, broadly construed as consistency of measurement.

On Messick's account, construct validity is the core evidential concept. In fact, construct validity is all there is to validity. Messick writes, "…the evidential basis of test interpretation is construct validity," (1989, p. 20) and that "…construct validity may ultimately be taken as the whole of validity in the final analysis," (p. 21). Later, he writes that

> construct validity, in essence, comprises the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and relationships with other variables (p. 34).

Construct validity is established through traditional validation procedures: by demonstrating the precision of tests, reliability (freedom from both systematic and unsystematic measurement error), appropriate exemplification of the construct. This latter criterion aims at minimizing construct underrepresentation and eliminating construct irrelevant test variance. Other kinds of evidence include convergent and discriminate evidence.

Convergent evidence purports to show that different facets of a construct correlate as stipulated by the theory of the construct under investigation, e.g., verbal intelligence and spatial intelligence as facets of general intelligence. Convergent evidence purports to support the claim that the different facets are in fact facets of the same construct. Discriminant evidence purports to show that the facets are not related to some other construct that could account for the observed correlations between facets. Messick, like Borsboom, acknowledges the role of formulating causal hypotheses about the processes underlying item response, but he is not so optimistic as Borsboom. Messick, unlike Borsboom, does not require that tests be constructed against the background of a causal model relating the construct to test performance. The role of causal

modeling is not antecedent to test development. Causal modeling is employed after a test is accepted.

The particular methods of amassing evidence in favor of construct validity are unimportant since Messick never tells us how much and, specifically, what kind of evidence justifies saying that an interpretive inference is valid. This is not surprising since he never tells us what it means to say that an interpretation is valid in the first place. We get a laundry list of different methods, most of which are correlational in character, and occasionally Messick gives admonitions such as "method X is never sufficient to establish construct validity" (*cf.* p. 35), but we are never told what *is* sufficient. This is unfortunate since it is the key to understanding Messick's account. If we were told what would be sufficient to establish construct validity, we would have an answer not only to the question of when attributions of validity are warranted, but we'd have an answer to the question of what it means to say that an interpretive inference is valid. Attributions of validity are warranted when the interpretive inference is valid, and since validity refers to a transparent epistemic situation (i.e., we know what evidence we have on the books), all we would need is to look at the record of evidence, decide based on some criteria (that Messick doesn't provide) whether the interpretive inference is sufficiently warranted to be deemed 'valid' and if it is, the attribution is warranted. But as things stand, we are left in the dark.

The situation is worse for Messick than it is for Borsboom who is also unilluminating when it comes to specifying conditions for making warranted attributions of validity. The reason for this is that for Messick, but not Borsboom, such conditions have a central and defining role in the theory. It is worth noting that fixing this problem with Messick's theory would not necessarily have ramifications for scientific realism (unlike with Borsboom). Not all interpretations claim or require the existence of attributes, whereas with Borsboom, the only relevant interpretation is one that requires commitment to the existence of attributes.

1.43 Pragmatic Concerns

Why do these epistemological concerns matter to the practicing psychologist? Often the relevance of epistemology to actual science is unclear, but hopefully this discussion of the epistemology of validity has made a persuasive case for the pertinence of epistemological concerns to the problem of validity. Both Borsboom and Messick give minimal requirements for justified attributions of validity. Borsboom requires causal analyses of item responses; Messick requires that constructs be representative, minimally contaminated by construct irrelevant variance, and that constructs be supported by discriminate and convergent evidence. However, more is needed. For example, it seems that not just any causal hypothesis would be adequate for Borsboom's needs. The causal hypothesis itself must be warranted independently of its ability to explain test behavior. Moreover, the requirement of a causal hypothesis is only a necessary condition for justifying attributions of validity. For his account to be of any use to the practicing psychologist more work needs to be done, e.g., sufficient conditions need to be provided. Messick's account, because it provides no guidance with respect to the question of when a psychologist can claim that his inferences are valid, is severely limited in utility. His account is silent on what it means for an interpretive inference to be valid. Second, his "theory of validity", with its laundry list of validation procedures, offers much in the way of methodological measures (purportedly relevant to establishing validity) but little in the way of normative requirements for attributions of validity. The practicing psychologist is left with a hodgepodge of validation procedures with no clear aim. Third, the practicing psychologist is given no indication as to how to assign meaningfully and non-arbitrarily the appropriate degree of validity to an interpretive inference. Messick tells us that validity comes in degrees, but he neither argues for this claim nor does he give any indication of the relevant evidentiary contribution of different kinds of evidence. What we are left with is an unworkable conception of validity. With Borsboom the aim is clear. Whether it is attainable will depend upon the question of whether scientific realism is justified in the context of psychometrics.

1.5 Conclusion

Neither of the two concepts of validity examined are beyond reproach, however I have argued that Borsboom's conservative conception of validity as a property of tests is preferable to Messick's conception of validity as property of interpretive inferences. On Borsboom's account, attributions of validity carry an ontological burden. Those who favor Messick's analysis need not be ontologically encumbered, though metaphysical frugality comes at the cost of the explanatory richness that only realist explanations afford. With a concept of validity in hand, the natural follow-up question is "how do we know if a test (or interpretive inference) is valid?" Borsboom gives a sketch of how one is to justify attributions of validity to tests, but since his main concerns are semantic and metaphysical, not methodological or epistemological, details remain outstanding. This leaves the philosopher of science and methodologist with the task of specifying when one can justifiably say that an attribute exists and that it produces variations in test scores. Messick, on the other hand, gives an account of validity rich in methodological options for justifying interpretive inferences, but, like Borsboom, says little about the requirements for validity. I have shown that for Borsboom's account this omission is benign, but for Messick's account the omission is pernicious.

2. Scientific Realism

2.1 Introduction

Traditionally construed, scientific realism claims that theoretical terms postulated by successful science refer. There are many different formulations of the realist thesis. What they tend to share is commitment to the existence of at least some theoretical posits of successful theories or the possibility of evidence accruing in favor of hypotheses that posit theoretical entities such as electrons, quarks, or psychological attributes. For some realists, it is the success of theories that warrants the ontological commitment to theoretical entities. Antirealism, on the other hand,

limiting credulity to the mere utility or empirical adequacy of theoretical posits, denies that success—explanatory, predictive, or otherwise—warrants commitment to the existence of theoretical entities. I will not argue for the truth of realism or antirealism generally. Following a general discussion of the semantic, metaphysical, and epistemological components of realism, I will offer a concise explication of two species of scientific realism that connect well with issues in psychometrics, Ian Hacking's *entity realism* as formulated in his (1983) *Representing and Intervening* and Jarrett Leplin's *minimal epistemic realism* as formulated in his (1997) *A Novel Defense of Scientific Realism* and subsequent articles. The tripartite approach to explicating realism adopted here has precedent in Horwich (2004) and Leplin (2005).

The relevance of scientific realism to the current discussion of validity cannot be understated. For example, the natural question to ask when presented with Borsboom's analysis is "what does it mean to say that a psychological attribute exists?" Or, perhaps more importantly, "when are we justified in saying that a psychological attribute exists?" Both of these questions are implicated in debates over scientific realism. Thus far I've argued that Borsboom offers little in the way of answering the second question. His account is foremost a metaphysical account and does little to address epistemological queries such as when one is epistemically justified in saying of a test that it is valid. I've also argued that Messick's account, too, offers little guidance in answering this question. However, both Messick and Borsboom claim to be realists. Since Messick's account of validity does not commit him to the possibility that evidence can accrue in favor of claims regarding the existence of psychological attributes, his account need not be concerned with answering the aforementioned two questions. Borsboom's account, on the other hand, faces serious threats if sense cannot be made of the existence of psychological attributes since it is a core thesis in his account of validity that psychological attributes exist. I've suggested that it is unfortunate, though not a damning feature, of Borsboom's account that he does not address the epistemological question; however, there may be insight to be gleaned from the philosophy of science and discussions of realism.

2.12 The Semantic Component of Scientific Realism

The semantic component of scientific realism, henceforth 'semantic realism', claims that propositions about theoretical entities should be taken literally and should be regarded as having determinate truth-values, either 'true' or 'false'. The truth of propositions concerning theoretical entities is independent of scientific methods, measurement procedures, and the ability to detect features of the world that resist direct observation. Truth-makers for propositions containing theoretical terms are facts about the structure of the world. That is 'There is an electron' is true if and only if there is an electron; 'electron' refers if and only if there is an electron.

Semantic antirealism comes in many forms. Consider verificationism, the philosophy of science according to which claims about theoretical entities are not assertions about the unobservable world (as the semantic realist would have them be); they are, rather, claims about observable phenomena. For example, the proposition 'There is an electron', according to the verificationist, is not a claim about entities. 'There is an electron' has no truth-value independent of a method of verification and correspondence-rules that cash out the meaning of the proposition in terms of observable phenomena. The truth-makers for propositions containing theoretical terms are facts concerning observable phenomena, i.e., not facts about unobservable entities, and the content of such propositions is exhausted by the outcomes of measurement procedures. Note that there are many ways to be an antirealist and that denying semantic realism is one way, but it is not the only way. Furthermore, one can be an antirealist and yet adhere to semantic realism *à la* Bas van Fraassen (1980).

2.13 The Metaphysical Component of Scientific Realism

Referential success of theoretical terms does not guarantee descriptive success. Theoretical claims may succeed in referring to unobservable entities, but that does not entail the descriptive success of theories and this is where the metaphysical commitments of realism outstrip the semantic

commitments. In other words, securing reference does not entail having an (approximately) true theoretical description of the entities that theoretical claims denote.

2.14 The Epistemic Component of Scientific Realism

The epistemic component of scientific realism, henceforth 'epistemic realism', is the commitment to the possibility of evidence accruing in favor of theoretical beliefs. Some realisms go further and claim that it is not only possible for evidence to justify belief in some theoretical claim or other, but that there are actual cases in which a set of epistemic standards for justifiably believing the claims entailed by some theory are satisfied; therefore, we are justified in believing some theoretical claim or other of some scientific theory. But going this far is not necessary for one to be a scientific realist. Claiming that evidence can *ever* warrant theoretical beliefs already puts one at odds with some of the most notable antirealists. What distinguishes some realists from others is what they take to be the conditions for warranting theoretical beliefs. They may also differ in what is taken to be the proper recipients of our credence. For example, some realists believe that theoretical entities (and not theories) are proper objects of credence, whereas others endorse no such restriction, i.e., they believe that we can be warranted in attributing (some measure of) truth to theories as well as existence claims for theoretical entities. In what follows, I will contrast two versions of scientific realism.

2.2 Entity Realism

The first version of realism that I will discuss is Ian Hacking's *entity realism*. I focus on this specific variety of realism because elsewhere Borsboom has argued that entity realism is required to make sense of certain methodological decisions in psychometrics (Borsboom, 2005). Borsboom explicitly embraces entity realism as the appropriate philosophy of science for latent variables in psychometrics. Also, entity realism is congenial to Borsboom's account of validity

since the psychological attributes that are purportedly measured by valid tests are usually intimately tied to latent variables such as general intelligence or extroversion.

Hacking (1983, p. 21) characterizes scientific realism as the position according to which "the entities, states and processes described by correct theories really do exist".[32] This is a fairly rough formulation of the realist thesis, and in fact the realism espoused by Hacking is slightly more refined. Being a realist usually entails some commitment to the truth of successful scientific theories (sometimes called 'theory realism') or to the ontological claim that some theoretical entities exist. Hacking opts for the latter, but not the former. He believes that we are warranted in believing in the existence of those theoretical entities scientists exploit in their investigations of "other *more hypothetical* parts of nature" (p. 265).[33] We may call this 'the more-H condition'. Hacking's notion of a theoretical entity includes, but is not limited to, "particles, fields, processes, structures, states and the like" (p. 26). If an entity can be manipulated or used as an investigatory tool, then it is real and, consequently, we are justified in believing that it is real. Thus, we have a putative epistemic criterion to accompany Borsboom's account which may answer the second question "when are we justified in believing that a psychological attribute exists?" An additional attractive feature of Hacking's account is that it allows for commitment to particular theoretical entities without being committed to any particular theory. So, we may be committed to, say, general intelligence without being committed to any particular theory of general intelligence. But, as usual, things are not so simple.

There is a general objection to Hacking's position worth mentioning. J.D. Trout (1999) notes that devices used to manipulate electrons and positrons are designed on the basis of theories that tell us what such theoretical entities are like. Without background theories about theoretical entities, we cannot know that the entities we take ourselves to be manipulating are in fact the ones being manipulated. We may be convinced to be realists regarding some entity on basis of some

[32] All quotations of Hacking are from his (1983) *Representing and Intervening*.
[33] My emphasis.

experimental manipulation, but prior theory makes such manipulation technologically possible. Experiment is indebted to theory whether or not we appreciate it when we are convinced by experimental results. Theory guides manipulation. General objections, such as Trout's, are instructive and worth keeping in mind; however, I will focus my attention on particular challenges to entity realism posed by psychometrics.

It is unclear how Hacking's position could be applied in the domain of psychometrics, since it is unclear how the more-H condition applies. What would it be to manipulate a psychological attribute in order to investigate more hypothetical or lower-level psychological phenomena? An analogy from psychopharmacology suggests itself since, on the face of it, psychiatric pharmaceuticals are prescribed to affect psychological attributes; therefore, in the case of psychopharmacology we have a *prima facie* case of intervention with respect to the properties or behavioral dispositions in question. Thus it makes sense to look at such cases in hopes of elucidating how a psychological attribute may satisfy the manipulability criterion. Many psychiatric pharmaceuticals are prescribed for their ameliorative effects on psychological disorders without there being a consensus on how these drugs work. Such is the case with Wellbutrin (bupropion) and the mechanisms by which it affects depression, addiction, and ADHD. The mechanisms by which other pharmaceuticals affect depression, such as selective serotonin reuptake inhibitors (SSRIs), e.g., Prozac (fluoxetine), are better known. The entity realist might say that the manipulation of an attribute such as depression by means of psychiatric intervention (by means of Wellbutrin, SSRIs, or monoamine oxidase inhibitors (MAOIs)) enables us to investigate more hypothetical neurological bases of the attribute. But this approach seems to hold little promise. I suspect that it may be circular. Unless one holds that depression is something over and above the neurochemical phenomena that give rise to it, the manipulation of depression by psychiatric intervention aimed at investigating the neurological basis of depression looks a lot like manipulating depression to investigate depression. Depression is not more hypothetical than itself, therefore such an investigation could not warrant belief in the existence

of depression. It is unclear that high-level psychological attributes can play a role analogous to Hacking's electrons. Consider general intelligence in particular. It is never manipulated in experiments nor does it seem to fulfill the role of an investigatory tool in the more-H condition. This is not to say that such a manipulation is not possible. Once can imagine experiments where variations in altitude or blood alcohol content might effect changes in general intelligence, but I know of no such studies nor is it clear how in such experiments we would be exploiting general intelligence to investigate more hypothetical phenomena.

Now let us consider the claim that we can be committed to the existence of some theoretical entity *sans* commitment to a theory that posits it. The fact of historical inconstancy of theoretical commitment recommends modesty with respect to the epistemic standing of theory. Unconditional faithfulness to theory is not only epistemically immodest, it is dogmatic, for it prohibits revisions to theory demanded by new evidence. As the history of science teaches us, the relationship between theory and evidence is not always favorable and we should be open to shifting our theoretical commitments in light of data. The foremost attraction of entity realism is its responsiveness to the difficulty posed by the fact that theoretical descriptions (or "stereotypes") of entities are subject to revision. Hacking claims that it is:

> Stereotypes may change as we find out more about a certain kind of thing or stuff. If we do have a genuine natural kind term, the reference of the term will remain the same, even though stereotypical opinions of the kind may change….Still, [van Fraassen] does tease the realist, who is confident that there are electrons: 'Whose electron did Millikan observe; Lorentz's, Rutherford's, Bohr's or Schrödinger's? (*The Scientific Image*, p. 214). Putnam's account of reference provides the realist with the obvious reply: Millikan measured the charge on the electron. Lorentz, Rutherford, Bohr, Schrödinger and Millikan were all talking about electrons. They had different theories of electrons. Different stereotypes of electrons have been in vogue but it is the reference that fixes the sameness of what we are talking about (pp. 80-81).

Therefore, reference for 'electron' is robust with respect to theory change. Commitment to the existence of the electron is justified independent of any particular theoretical description of it or "common core" of a plurality of theoretical descriptions (*cf*. p. 264).

Unfortunately, psychological attributes such as general intelligence cannot be stripped of theory in the way that Hacking believes electrons can. There are at least two ways of arguing for this claim. One could argue, as Trout does, that Hacking is wrong with respect to theoretical entities in general: for no theoretical posit can we be justified in believing that it exists without also being committed to some theoretical description of it. Psychological attributes come out the same as electrons in this case. Ontological commitment does not come unburdened by theory. This general strategy, if successful, would satisfy the aim of the second strategy. The second strategy would be to focus on psychological attributes in particular and show that relevant ontological commitments do not come "theory-free", so to speak. Either strategy, if successful, serves to establish my claim.

There are specific reasons to doubt that the commitment to psychological attributes is separable from psychological theory. The difficulty arises when we consider the theory of meaning to which Hacking subscribes, which was developed by Hilary Putnam (1979). Take the referring term 'dog'. Putnam analyzes its meaning along four parameters: syntactic markers, semantic markers, stereotype, and extension:

> *dog*: [concrete, count noun], [mammal, names a natural kind],[has four legs, is a domesticated animal, sometimes used for service work or other utilitarian purposes, lives on land],[…]

Syntactic markers indicate grammatical features of the term and whether the term is concrete or abstract. Semantic markers designate categories under which the term is correctly applied. Stereotypes list commonly held beliefs about the things to which the term applies. Stereotypes may be turn out to be false or in need of refinement. For example, if in the future PETA were to push through legislation that banned owning dogs as pets or using them as service animals, it would not be a contradiction to speak of my pet Doberman Sebastian or search and rescue dogs in WWII. Stereotypes enable successful use of terms within a linguistic community. The extension of 'dog' just is the set of objects denoted by the referring term.

It is not difficult to see how this scheme applies to electrons, but how does it work for, say general intelligence? First of all, no psychometrician has any idea what the extension of 'general intelligence' is exactly. It is likely that it is a multiply realizable intellectual capacity, but pointing to any one of its realizers only partly pins down the extension of the terms. Unlike money there is no general description of general intelligence that enables us to recognize when it is being realized. There is but one way to tell whether general intelligence is being realized which is to perform a factor analysis on the task in question to see if it loads on the *g*-factor, but to do this is to land oneself smack in the middle of theory. To say that the realizers of general intelligence are those things that can be ranked according to their performance on items or tests that load on the *g*-factor is to embrace certain theoretical claims regarding general intelligence and measurement theory. The extension and the stereotypes are not cleavable. One may object that this characterization is unfair, that electrons are *things*, but general intelligence is not a thing; it is, rather, a property. I take it that properties, magnitudes, capacities, etc. also can be theoretical entities (as does Hacking), possibly subject to manipulation, so I do not think the difference between objects and properties is important for this discussion. After all, it is not electrons that we measure, it is properties of electrons and the measurement of these properties supposedly warrants the postulation of electrons.

Hacking's position in *Representing and Intervening* cannot be applied to psychological attributes in a direct or obvious way. His entity realism seems to have been constructed with physics (and perhaps biology) in mind, but not the behavioral sciences. In fact Hacking expresses skepticism regarding realism with respect to psychometric constructs central to this dissertation:

> We can measure IQ and boast that a dozen different techniques give the same stable array of numbers, but we have not the slightest causal understanding. In a recent polemic Stephen Jay Gould speaks of the 'fallacy of reification' in the history of IQ: I agree (p. 39).

Nevertheless something seems correct about the idea that manipulability warrants ontological commitment. It seems that psychometricians, such as Arthur Jensen, have something akin to Hacking's entity realism in mind when they claim that the vulnerability of general intelligence to inbreeding depression gives us reason to believe that general intelligence is a real. Inbreeding depression, when evident, provides the context of a natural experiment in which we can discern the manipulation of IQ in groups over time. Stereotype threat scenarios (Steele & Aronson, 1995), too, give reason to believe that IQ can be manipulated. In such scenarios the IQ scores of minority students were significantly lower when they were told that they were taking an intelligence test than they would have been had they not been told they were taking such a test. The scores of white students did not show an appreciable difference in performance when told that they were taking an intelligence test. Contrary to the skepticism expressed in the quotation above, we *can* manipulate IQ (such as in stereotype threat scenarios) and, therefore, we *do* have some meager causal understanding of IQ. For this reason, entity realism is relevant to the discussion of realism in psychometrics and its mark on "psychometric realism" will be obvious.

2.3 Minimal Epistemic Realism

Given the apparent inextricability of psychological attributes and substantive theory, a more robust realism than entity realism, one that is not prejudiced against theory, is required. More recently, Jarrett Leplin has formulated a nuanced statement of scientific realism:

> Theoretical entities that are needed to explain or predict empirical results, and that are posited by well-supported theories free of empirical or conceptual difficulties, exist and have those of the properties these theories attribute to them that enable them to fulfill their explanatory and predictive roles (p. 5).

Minimal epistemic realism claims that there are empirically realizable conditions (to be specified below) such that were they to obtain, we would be justified in taking a realist stance toward a theory and certain of its posits. Leplin adopts an explanationist defense of realism: were realism

not true, explanatory and predictive success would be inexplicable. However, Leplin is specific about the kinds of explanatory and predictive success as well as the entities posited by theories that require a realist interpretation. Predictive novelty is sufficient for justifying realism. Novel predictions in particular are best explained by realism with respect to theory. For a result to be novel for a theory, two conditions must be satisfied, an independence condition and a uniqueness condition.

Independence of a result R with respect to a theory T that predicts R requires that R was not essential to the construction of T. The motivation for this condition is to block objections that T was designed to predict R. If T were constructed to predict R, then that fact alone would explain T's predicting R, realism would be unnecessary.

Uniqueness of a result R with respect to a theory T that predicts R requires that no extant viable alternative to T also predicts R. The motivation for this condition is to block objections that a result R may be independent with respect to two incompatible theories each of which predicts R. Since the population of theories contending to explain some set of phenomenon is subject to change, ascriptions of novelty are indexed to a time (or else they would be variable).

With these two conditions in hand, Leplin offers the following characterization of novelty:

> A result R is *novel* with respect to a theory T if it satisfies the independence condition with respect to T and, at the time that T first predicts it successfully, no viable rival to T also predicts it (p. 11).

Those results that satisfy the uniqueness and independence conditions with respect to a theory are the ones that justify belief in theoretical hypotheses. Realism is invoked to explain a theory's sustained record of novel predictive success. Leplin provides the following historical examples of successful novel predictions:

> the use of Newtonian theory to discover the outer planets of the Solar System,…, Mendel's backcross test of the genetic hypothesis,…, the application of atomic theory to

> Brownian motion,…,the bright spot discovered in spherical diffraction,…, the conversion of matter and energy,…[and] the gravitational deflection of starlight, (Leplin 2004, p. 129).

Leplin's brand of realism is itself a scientific theory. It makes predictions and is therefore subject to empirical test; it predicts that theories enjoying novel predictive success will continue to enjoy success, entities postulated by these theories will maintain the properties ascribed to them by theory, and viable empirically successful rivals to epistemically justified theories will not arise. Since these predictions satisfy the uniqueness and independence conditions, they would, if true, count as novel for realism, and thus realism would be epistemically justified by its own lights (Leplin, 2004).

An immediate problem is that in psychology we having nothing like the mature theories that Leplin cites. Psychological attributes are mired in theoretical commitments, but for any particular attribute there is nothing approaching a comprehensive fleshed out theory from which we could deduce predictions of behavior. At most, it seems that we have minimal theoretical constraints imposed on psychological attributes from cognitive science and neuroscience. While *prima facie* there is promise in item response theory (IRT) in that IRT models enable the prediction of test performance from position on a latent variable, IRT models, arguably, afoul of the independence condition. The models are designed to predict the data.

2.4 The Antirealist Alternative

Realism's philosophical foil is antirealism. To be an antirealist is to deny any one of the three components of scientific realism given above, i.e., semantic realism, metaphysical realism, or epistemic realism. One need not deny all three components of realism to be considered an antirealist. Denying epistemic realism is typically enough. For example, Bas van Fraassen in his monumental (1980) *The Scientific Image* formulates an antirealist position known as "constructive empiricism" which denies both metaphysical and epistemic realism, but not

semantic realism. According to constructive empiricism, our epistemic limitations are such that we could never be warranted in making theoretical claims; theoretical knowledge is out of reach and the only structures about which we can have knowledge are observable structures and the empirical adequacy of theoretical claims. Additionally, evidence cannot warrant hypotheses that posit theoretical entities, which would include psychological attributes. Clearly the alternative to realism does not sit well with Borsboom's account. A semantic realist, van Fraassen argues that the epistemic resources necessary for justifiably claiming that a theory is (at least partially) true are unavailable.

Operationalism, a philosophy of science originally formulated by Nobel Prize winning physicist Percy Bridgman (1927), denies the semantic and metaphysical components of scientific realism. According to operationalism theoretical concepts such as intelligence have their content fully specified by a measurement procedure. Bridgman writes

> To find the length of an object, we have to perform certain physical operations. The concept of length is therefore fixed when the operations by which length is measured are fixed: that is, the concept of length involves as much as and nothing more than the set of operations by which length is determined. In general, we mean by any concept nothing more than a set of operations; *the concept is synonymous with the corresponding set of operations*, (Bridgeman, 1991, p. 59)

Operationalism is silent as to whether theoretical terms refer. The question 'do Xs exist' is abandoned in favor of the question 'is 'X' meaningful', and this latter question is answered by whether there is a set of operations that define the concept of an X. If there is no procedure for measuring general intelligence, then sentences containing the term 'general intelligence' are meaningless. Theoretical terms cannot be said to enjoy referential success independent of our measurement procedures or scientific methods. This entails a violation of semantic realism. Operationalism in psychology has fallen out of favor as an overarching philosophy of science, though its influence still persists in methodology. Psychologists and psychometricians are reluctant to admit into their catalogue of theoretical entities/properties constructs for which there

is no measurement procedure. For now I will set to the side the objections to operationalism and reasons for thinking that operationalism renders psychometric practice unintelligible. I will return to this point in the next chapter when I argue that realism alone can make sense of psychometric practice.

Conventionalism covers many antirealist positions, including operationalism; it refers to a class of positions characterized by a denial of metaphysical realism and semantic realism. Conventionalism (sometimes referred to as 'constructivism'), asserts that theoretical terms denote not properties or structures, but mental artifacts. 'Intelligence', 'electron', 'charge', and the like are simply categories that we place on the world and fail to refer to anything that exists independently of our measurement procedures or scientific methods. Theoretical terms connote useful fictions.[34]

3. Conclusion: Psychometric Realism

I have considered two different forms of realism as well as the antirealist alternative. It should be clear that if one is to embrace an antirealist alternative, then one cannot justifiably claim that a test is valid (in Borsboom's sense) provided one is also a semantic realist. If names of psychological attributes are taken at face value, as purporting to denote theoretical entities, then evidence cannot warrant the claim that a test measures that attribute. My selection of realist theses was not idiosyncratic. Both Hacking and Leplin formulate realist positions that are relevant to scientific realism in the context of psychometrics (and, thus, the question of test validity). Hacking's position seems at odds with realism about general intelligence at first glance, but the

[34] Borsboom *et al.* (2003) claim incorrectly that van Fraassen's constructive empiricism is a form of conventionalism. While constructive empiricism is at odds with metaphysical and epistemic realism, it requires semantic realism which conventionalism forbids. The source of this confusion is Borsboom *et al.*'s belief that to deny theory realism is to deny that theories can be either true or false; i.e., it is to conflate the rejection of epistemic realism with regards to theories with the rejection of semantic realism with regards to theories. van Fraassen rejects entity and theory realism; however, he does believe that theories have truth-values and that sentences containing theoretical terms should be interpreted literally. We just can't know what those truth-values are since we cannot amass evidence for theoretical claims.

spirit of entity realism is clearly in line with what motivates realist intuitions among psychometricians. Leplin's position offers robustness where Hacking's realism is anemic, namely with respect to epistemic attitudes towards theories. Leplin's position is also clear about the kinds of evidence that are relevant to justifying realism. For realism to be a viable philosophy of science for psychometricians, it will need to capture the spirit of Hacking's position in order to make sense of their realist intuitions. It must also bring within its purview realism about theories. That is, the proponent of *psychometric realism* needs to be a realist about both entities and theories. Psychometric realism will be the topic of the next chapter. I will then argue that psychometric realism is required to make psychometric practice intelligible, especially in the field of intelligence research, for realist presuppositions are pervasive therein. I will then examine the question of whether psychometric realism is justified with respect to extant theories of intelligence whose central theoretical posit is general intelligence, i.e., "g-factor" theories of intelligence. Therefore, I leave open the possibility that psychometric realism may be the only philosophy of science that can make sense of psychometric research on intelligence, but that no theory of psychometric intelligence is (or can be) warranted by evidence.

CHAPTER THREE

PSYCHOLOGICAL MEASUREMENT, METHODOLOGICAL REALISM, AND PATHOLOGICAL SCIENCE

We milk the cow of the world, and as we do
We whisper in her ear, 'You are not true.'

Wilbur (1957) from *Epistemology*

1. Rationality in Psychometrics

Psychometrics is the branch of psychology devoted to the study of individual differences and psychological and educational assessment. At its inception in the early 20th century the field was unabashedly realist in its orientation. Constructs identified through statistical analysis were hypothesized to refer to "real" psychological attributes. For example, Spearman and others took the fact that tests of intelligence all correlate positively to indicate that there was a latent common cause of the pattern of intercorrelations. This latent common cause manifested itself statistically

as a general factor that accounted for (i.e., screened off) a certain portion of the correlations between measures of mental ability. Spearman hypothesized that this general factor referred to a property of the mind, namely "general intelligence," (1904, 1927). With the ascendancy of logical positivism, psychometrics assumed a different philosophical orientation. Operationalism took hold and in psychometrics, if not in psychology more broadly, constructs were defined in terms of measurement instruments. For example, Boring, when asked to define 'intelligence', famously claimed "Intelligence is what intelligence tests test," (Boring, 1923). The concept of psychometric validity, once cashed out in terms of whether a test measures the attribute it purports to measure (Kelley, 1927), later was formulated in terms of whether performance correlated with some other criterion such as academic or occupational success, and likewise for psychological constructs: they too were defined purely in terms of their statistical relationships with observable criteria such as academic achievement or occupational success. Hence, attributes such as "general intelligence" were stripped of their psychological significance and defined in terms of their statistical relationships with other measures (Cronbach & Meehl, 1955). Under this scheme, general intelligence would be that thing which correlates such and such with occupational success, not (hypothesized to be) mental energy *a la* Spearman.

Recently however, there seems to be a shift back to a more "realist" psychometrics, or so I will argue. There are obvious examples of this turn (*cf.* Borsboom, 2005; Borsboom, Mellenbergh, & van Heerden, 2003), but I wish to argue for something more than that there are some psychometricians who openly endorse realism in some form. I will argue that the rationality of methodological decisions and practices implicated in theory formation in psychometrics is undermined unless we attribute at least minimal epistemic realism to psychometricians, regardless of the philosophical positions that particular psychometricians openly espouse. Minimal epistemic realism is the thesis that it is, in principle, possible to justify claims about unobservable entities. I focus on methodology as implicated in theorizing to emphasize the role of realism in the *development* of theories and not theory adjudication itself. It is the latter I take to be the object

of scrutiny for those philosophers of science impressed by underdetermination theses—those who argue that theory adjudication cannot be grounded solely in empirical data, and that, therefore, theory choice must appeal to non-epistemic virtues. The unobservable entities in question here are latent traits, those entities to which latent variables purport to refer. The truth of minimal epistemic realism is not under dispute here; what is under dispute is whether minimal epistemic realism is a necessary presupposition for the rationality of at least some scientific decision making in psychometrics. I take up the former dispute elsewhere. Methodological realism, the position defended here, is simply the claim that there are important parts of psychometric methodology that cannot be counted as rational unless minimal epistemic realism is presupposed.[35]

Methodological realism and minimal epistemic realism are logically independent; therefore, one may assent to one while denying the other. Similarly for minimal epistemic realism and ontological realism in psychometrics, which is the thesis that the latent traits or attributes that figure in psychometric inquiry exist.[36] Since methodological realism may be true even if ontological realism is false, they too are logically independent; therefore, vindicating methodological realism does not license attributing presumptions of ontological realism. That is, the rationality of certain methodological decisions requires a commitment to the possibility of evidence accruing in favor of hypotheses that posit unobservable entities, and yet those entities may not exist. Methodological realism's tenability will be assessed with respect to methodological practice in the psychometric study of intelligence, though much of what I will have to say can be generalized to other areas of psychometric inquiry, such as personality testing, educational assessment, and sociological research (Chen *et al.*, 2006). Specifically, I will argue

[35] I borrow this term from Jarrett Leplin (1986) who argues for methodological realism in Millikan's discovery of the electron's charge.

[36] This thesis is akin to, but distinct from, the metaphysical realism often taken to be associated with scientific realism. Ontological realism in this case is a special case of metaphysical realism, the latter being associated with scientific realism generally; ontological realism, for the purposes of this paper, is restricted to particular theoretical posits.

that model selection for representing covariation in test performance (across tests) provides strong grounds for methodological realism, though not for ascribing presuppositions of ontological realism.

I will then turn to classical and modern test theory. Classical test theory claims that observed scores on tests are the product of two sources: "true score" and error. Thus variability in test performance is decomposed into variability in true score and error variance. Modern test theory is almost synonymous with item response theory which posits that performance on *items*, i.e., not tests, is probabilistically related to one's position on a latent variable. So the two differ with respect to the unit of analysis, item vs. test, and the role of probability: classical test theory is deterministic and modern test theory is not deterministic.[37] A strong *prima facie* case could be made that adopting item response theory as opposed to classical test theory reveals a presumption of ontological realism. For example, it seems that item response theory involves the kind of unwarranted reification of statistical factors that Stephen Jay Gould attacks in his *Mismeasure of Man* (1981, 1996). I will show that the situation is actually quite complicated, but that proponents of item response theory need not presuppose ontological realism in order to make sense of their selection of item response theory over classical test theory. Moreover, the claim that the selection warrants methodological realism is undermined by the manifold internal reasons for rejecting classical test theory. What does require epistemic *and* ontological realism, however, is the estimation of ability in the context of item response theory.

If methodological realism is true, and it will be argued that it is, then the fact that psychometricians typically proclaim agnosticism or antirealism obscures their (latent) epistemic aspirations, if their methodological decisions are nonarbitrary and sensible, i.e., rational. Additionally, the truth of methodological realism will require a reassessment of Stephen Jay Gould's criticisms in *The Mismeasure of Man*.

[37] I leave to the wayside how to interpret these probabilities. While it is certainly important in this context how the probabilities are interpreted (as in all contexts), the concern is tangential to the argument of this paper.

Foremost among Gould's grievances with psychometrics is the alleged ubiquitous tendency to reify latent variables, including *g*. Methodological realism in the context of psychometrics holds that realist presuppositions underpin psychometric research. And so it seems that Gould's criticisms may yet have a target in the face of psychometricians' professed epistemic and ontological modesty. An interesting twist on the debate is that Gould's objections, dismissed by psychometricians as being outdated (even at the time they were originally published), could prove cogent after all, though not for the reasons he gives. Psychometricians are correct in their response to Gould's accusations that they overtly and promiscuously reify latent variables (as latent traits). They (*cf.* Carroll, 1995; Bartholomew, 2004) counter Gould by citing their own ontological commitments and epistemic goals. In chapter 1 of the dissertation I argued that it is doubtful that Gould's charges hold even for Charles Spearman and Arthur Jensen, two psychometricians to whom Gould devotes much effort to portray as pathological "reifiers." However, methodological realism possibly offers Gould an opportunity to refocus his objection. The Gouldian argument against psychometrics shifts from manifest ontological commitments to the underlying epistemic commitments guiding practice. Using methodological realism as a toehold, the Gouldian may find a position from which to attack minimal epistemic realism (with respect to psychometric theoretical entities). I will assess this Gould-inspired objection and argue that methodological realism in the context of psychometrics is unobjectionable.

Joel Michell (1997, 1999, 2000, 2004, forthcoming) has argued along Gouldian lines claiming that psychometrics is an instance of "pathological science." The basis for this critical diagnosis is that psychometricians, without evidence, assume that psychological attributes have a quantitative structure. Thus, Michell's critique would seem to presuppose that those who are doing pathological science, at least in this case, have realist commitments. Michell's complaint can be understood as an admonishment to justify the theoretical commitments underlying psychometric practice. That is, like Gould, Michell attacks dogmatic ontological commitment to psychological attributes having certain theoretical characteristics (e.g., having a quantitative

structure). Further, according to Michell, psychologists and psychometricians not only uncritically assume the hypothesis that psychological traits are quantitative, but they also disguise this assumption to deflect criticism. I will also evaluate the claim that psychometrics is "pathological science." I will argue that Michell's criteria for what counts as an instance of pathological science indict much of normal science, that the notion of a pathological science is not clearly defined, and that to the extent that sense can be made of this notion, it is unclear whose burden it is to treat the pathology.

Before arguing that there are instances of psychometric practice that are plausibly construed as realist, either epistemic or ontological, I should note the obvious to preempt a potential objection. There need not be any tension between agnosticism and methodological realism. It is consistent to be agnostic about a particular latent traits while also being committed to the possibility that evidence could make one a believer (or nonbeliever); this latter commitment is tantamount to renouncing dogmatic agnosticism (i.e., skepticism) and embracing fallibilism. My argument is not that the agnostics are inconsistent; indeed, they may be quite prudent. However, if one holds that we can understand theorizing in psychometrics without realist commitments (e.g., either from a constructivist, instrumentalist, or empiricist standpoint), then we have a problem, for I argue that there are practices in psychometrics, including the process of theorizing, which require a commitment to at least minimal epistemic realism if not full blown ontological realism. Therefore, my arguments are directed at those who claim that latent variables are merely mathematical transformations and do not denote, nor should they be taken as possibly denoting, psychological attributes or traits.

2. Methodological Realism in Psychometrics

In this section I will argue that certain psychometric methodological decisions, stripped of realist commitments, seem to have no rational basis. Some cases will require stronger commitments than others, and the justification for these commitments will not be considered here. Rather, my

concern will be simply to argue that these commitments figure essentially in methodology. One kind of realist commitment is an epistemic commitment to the possibility of justifying claims that make reference to theoretical entities. The stronger commitment is a metaphysical commitment to the actual existence of some theoretical entity.

2.1 Model Selection Requires Epistemic but not Ontological Realism

What I find to be the most compelling piece of evidence in favor of methodological realism (in psychometrics) is articulated by Borsboom (2005). Borsboom argues that the preference for certain kinds of measurement models in psychology is telling of underlying realist commitments among psychologists and psychometricians. Consider the different psychometric models for representing the structure of individual differences in Figure 3.1 below.

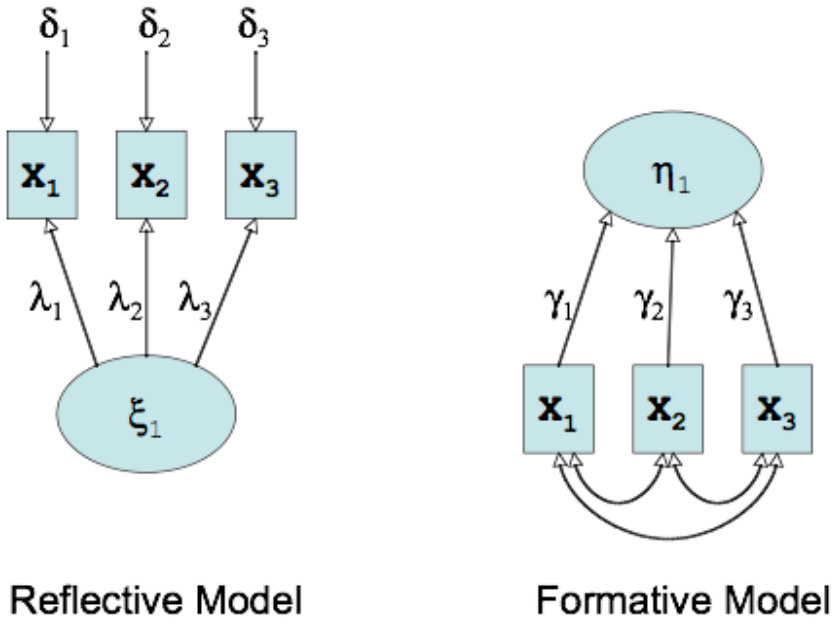

Figure 3.1: Reflective and formative measurement models.

Appropriating the terminology of Edwards and Bagozzi (2000) and Borsboom (2005), I will refer to the model on the left side of Figure 3.1 as a *reflective* model and to the model on the right as a

formative model. Each $\mathbf{X_i}$ is a manifest variable or indicator such as an item response or test variable. ξ and η are latent variables, each λ_i is the factor loading of each indicator in the left-hand model, and each γ_i is a weight of the indicator with respect to the latent variable. Each δ_i is an error term for the relevant indicator. The reflective model is the typical unidimensional measurement model found in psychology and psychometrics. In the measurement of general mental ability, each $\mathbf{X_i}$ would be, for example, a subtest (or item) of the Wechsler Intelligence Scale for Children (WISC) and performance on each subtest (or item) would be seen as a function of position on the latent variable *g*: it is differences in positions on *g* which cause differences in performance on the indicators, hence the direction of the arrows. Formative models are more popular in sociological research. For example, socioeconomic status (SES) is modeled with a formative model. In the formative model the direction of causal influence is reversed, running from the indicators to the latent variable. The latent variable is regressed on its indicators, not the other way around. One (or a population) occupies a position on SES because of the values of the indicators, such as gross yearly income or level of education, and SES is interpreted as summary of the observed measures; no ontological commitment is required. We may even use one's SES score to predict one's level on some unmeasured indicator, but even this does not entail that SES is being treated realistically. Such is not the case with latent variables in the reflective model.

Does the choice to model covariation with a reflective model, as opposed to a formative model, require realism? There is no *a priori* reason to prefer one model to the other unless one harbors realist commitments. Without such commitments, the choice is seemingly arbitrary. Indeed, it would seem irrational to prefer the model that carries with it ontological commitment if one does not embrace some form of realism. But now the question is what kind of realism is required? Borsboom believes that model selection of the kind under discussion requires both ontological realism and methodological realism in order to be non-arbitrary. That is, he believes that reflective model selection requires commitment to the existence of the latent trait as well as

commitment to the possibility of evidence accruing in favor of theoretical hypotheses that posit the latent trait. I will argue that the demands of rationality require only the latter.

It may be that many psychometricians commit themselves to the existence of latent traits; however, it does not seem that choosing to model covariation reflectively requires so much. Measurement models, such as those in Figure 3.1, are *hypotheses* and as such they may enjoy more or less empirical support. They are offered up as explanations of variations in test behavior in a population. Proposing a hypothesis does not itself commit one to the existence of the entities postulated therein. That psychometricians model covariation reflectively while remaining agnostic about the ontological status of their latent variables is perfectly intelligible; one might even argue that it is just good science to withhold commitment until sufficient evidence has accumulated that would warrant belief in latent traits. Psychologists may nevertheless attempt to justify their ontological realism by appealing to the model's ability to fit the data, or appeal to the explanatory or predictive power of a particular model to justify claims that posit latent traits, but there is nothing in the process of modeling or selecting a formative model that carries an ontological commitment.

While we can make sense of model selection without ascribing ontological realism to psychometricians, the rationality of model selection seems to require methodological realism. Since agnosticism with respect to latent traits is, at least nominally, the dominant position among psychometricians, their position would be aptly described as latent epistemic realism, or "latent realism" for short (with the understanding that latent realism does not entail latent ontological realism). Reflective models are hypotheses about the structure of covariation. They are also causal hypotheses about the underlying, i.e., latent, common cause(s) of test behavior. These models are selected not only for their structural and mathematical properties, but because they are testable against novel data through confirmatory factor analysis. A fit between the model and data is interpreted as confirmation of the model. A good example of such a study is Jensen and Weng's (1994) study where the authors sought to verify the presence of a dominant latent factor,

viz., the *g*-factor, in a variety of samples, both simulated and real, through exploratory factor analysis and confirmatory factor analysis. It would seem to make little sense to be "testing" these measurement models against data if Jensen did not think that there was the possibility of confirming the *g* hypothesis. Were his sole aim empirical adequacy and not the structure of mental ability, the selection of reflective measurement models would make little sense, as there are available formative models that also capture the data. Merely subjecting such models to a confirmatory factor analysis suggests that one is aiming for confirmation of the model, causality, unobservables, and all. Thus, the selection of models that posit latent sources of variation and the subsequent testing of those models, as opposed to models that do not make such empirical claims about the latent structure of ability, require methodological realism. Psychometric praxis is arbitrary unless we attribute to psychometricians (at least those who work on *g*) the belief that evidence can accrue favorably or unfavorably for the existence of latent traits.

2.2 Item Response Theory (IRT): A Complicated (but Instructive) Case

A related but distinct example of the role of methodological realism in psychometrics concerns IRT. IRT refers to a class of models used to predict test behavior on the basis of position on some latent variable and properties of test items such as difficulty, the item's capacity to discriminate, and the probability of guessing the solution to the item. IRT models also provide a means for estimating ability. IRT was developed in response to many problems that plagued classical test theory. An IRT model specifies an item characteristic curve (ICC), a monotonically increasing function that describes the relationship between position on a latent variable and test performance. That an ICC can describe performance is one postulate of IRT. The other is that performance can be predicted by position on the latent variable (Hambleton, Swaminathan, & Rogers, 1991).

There are further relevant assumptions of IRT models. First among them is unidimensionality. A set of items is unidimensional just in case the item(s) in that set measure

only one latent trait. In many IRT texts, including Hambleton *et al*.'s canonical introduction to the subject, unidimensionality is treated as an assumption of IRT models. It is arguable, however, whether unidimensionality is appropriately called "an assumption." For example, models are testable with respect to unidimensionality. If the posited latent trait can account for test performance, then the assumption is satisfied. If not, then it may be the case that the model needs to invoke more than one latent trait, in which case the appropriate model is multidimensional. Item response modelers routinely refer to unidimensionality as an assumption; the fact that the assumption is directly tested in fitting models, seems to suggest otherwise. Nothing in my argument hinges on the status of unidimensionality in this regard, so I'll leave this matter be. For the sake of simplicity, I will stick to unidimensional models.

The second assumption of IRT models is that of local independence. A model is locally independent when the correlation between performance on test items disappears when we conditionalize on the latent trait. In the parlance of the philosopher of science, the latent trait *screens off* the correlations between item responses much like the falling of a barometric pressure screens off the correlation between rain and falling barometers. Formally, we may express local independence with the following equation:

$$\Pr(R_1, R_2, ..., R_n \mid \theta) = \Pr(R_1 \mid \theta)\Pr(R_2 \mid \theta)...\Pr(R_n \mid \theta) = \prod_{i=n}^{n} \Pr(R_i \mid \theta)$$

where ' $\Pr$ ' denotes the probability function, ' R_k ' denotes a response on some item, and ' θ 'denotes a latent trait. Unidimensionality is sufficient, though not necessary, for local independence. Multidimensional models may be locally independent. What is necessary and sufficient for local independence is a specification of the complete latent space; a unidimensional model specifies the complete latent space if and only if that model also exhibits local independence.

An example of a one-parameter logistic IRT model is given by the following equation:

$$\Pr_i(\theta) = e^{(\theta - b_i)} / 1 + e^{(\theta - b_i)} \qquad i = 1, 2, 3, ..., n$$

where $\Pr_i(\theta)$, an S-shaped curve with value between 0 and 1, is the probability of answering item i correct given latent ability θ, b_i is the difficulty parameter of item i, *n* is the number of items, and *e* is a constant (2.718). In this case where *n*=4, this model defines the following ICC for four items, which is reprinted from Hambleton *et al*. (1991, 14):

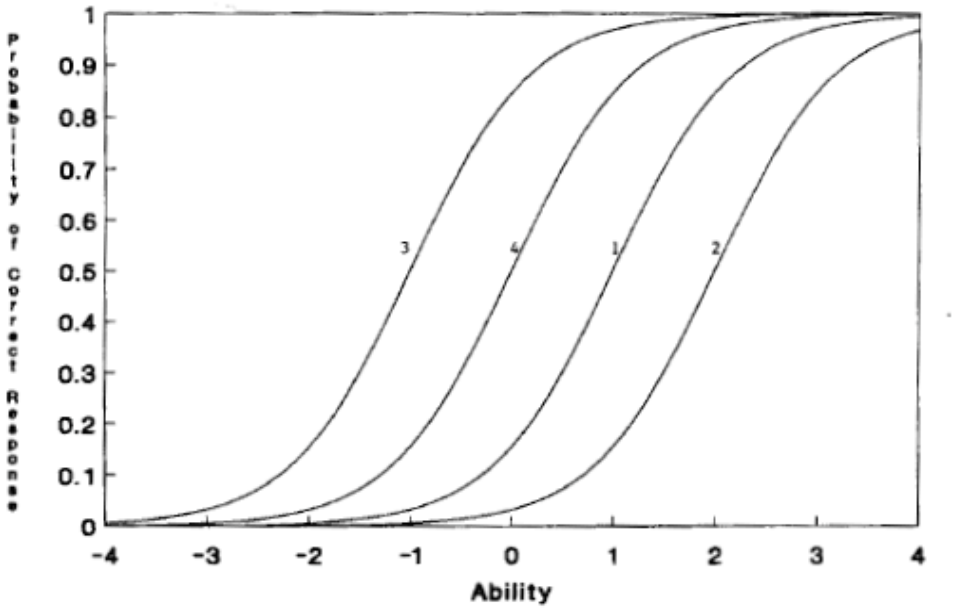

Figure 3.2: Item characteristic curves for four items as modeled by a one-parameter logistic IRT model.

Additional parameters could be added, such as a discrimination parameter, or a guessing parameter (also called a "pseudo-chance"), depending on the data to which one is fitting the model; however, for the present purposes (thinking about the status of θ), the above example is sufficient. Figure 3.2 demonstrates how the probability of correctly responding to an item is a function of the level of θ.

The question is this: is there something about IRT models and their place in psychometric theorizing that would commit psychometricians to ontological or minimal epistemic realism? At first glance, the answer seems to be 'yes' on both counts. Ontological realism is required by the fact that performance is regressed on latent ability θ. Minimal epistemic realism is required by the fact that models are subject to testing, and, therefore, the possibility of evidence accruing in favor of models that posit latent traits. Upon closer examination, however, things are not so simple.

2.21 When Realism is *not* Required: at Least One Use of IRT Does Not Require Ontological Realism

Ontological realism is *not* required for the use of IRT models to be rational. As in the case of reflective models, IRT models may be supposed to be hypotheses that make reference to unobservable entities. Conjecture need not carry commitment to the existence of the thing conjectured even if it does entail a commitment to the possibility that the thing conjectured exists. It is only acceptance of the model as true that would require metaphysical commitment to the existence of the latent trait, and only if 'θ' is read as denoting latent ability. If one reads 'θ' as simply a placeholder for confounding causal factors yet to be disentangled, commitment to latent ability may be eschewed. Even acceptance of the model does not seem to command commitment to the existence of a trait to which 'θ' refers. Following van Fraassen (1980), the acceptance of IRT models would seem to require nothing more than belief that the model is empirically adequate. There is no obstacle to treating IRT models as merely useful predictors of test behavior. This conclusion will seem, if not patently false, strange to psychometricians and proponents of IRT, or anyone who has read an elementary textbook on IRT models since it is typically listed among the assumptions of IRT models that there are underlying factors that explain test performance and these factors are the latent factors that appear in IRT models, i.e., θ (*cf.*

Hambleton, *et. al*, 1991). I am arguing that such an assumption is metaphysically ostentatious and unnecessary. Since the deployment of IRT models does not require such a commitment, the commitment is not an assumption of the models. I am not arguing that such commitment is never warranted, only that it is not required to make sense of using IRT models to predict item response, even if the relevant test items are IQ test items. And while I may seem to be pitting myself against psychometric dicta, in fact I am exonerating psychometricians (or at least offering them an escape) from charges that their methods require the unreflective reification of latent factors.

2.22 Opting for IRT as Opposed to Classical Test Theory Also Does not Require Epistemic Realism.

Now let us consider matters from an epistemic side. Are proponents of IRT implicitly committed to the possibility of accruing evidence in favor of latent traits? After all, by using IRT one is indicating a preference for IRT over classical test theory, just as the intelligence researcher fits reflective models to data. Certainly evidence can accrue in favor of a particular model; when the model fits novel data, psychometricians take that as confirmation of the model. But must such tests be interpreted as being in part founded on a presupposition that evidence can accrue in favor of (or if the model doesn't fit, against) latent traits? The answer: it depends.

What makes the case of preferring IRT to classical test theory any different from the selection of reflective over formative models? Could we not make parallel arguments against the claim that selecting reflective models requires latent realism? The answer: No. In the case of model selection, it was not the use of the model *per se* that was our *explanandum*, rather it was the choice of a model that contains causally efficacious latent traits over a model where the latent trait is not causally efficacious and is plausibly interpreted as a weighted sum score of measures on certain manifest indicators. Then once that choice is made, a particular reflective model is subject to confirmatory tests. In the case of IRT, however, we are not considering a choice

between a test theory that does posit causally efficacious latent traits and a statistically equivalent theory that does not; if such were the case and IRT were the dominant test theory, those facts would require a realist explanation just as the case with the covariation models in section 2.1. The two cases would be strictly analogous. One might object that such a selection *is* going on with IRT. When we use IRT we are implicitly choosing it as a test theory over classical test theory (CTT), and since the test theory we are selecting makes reference to latent traits, by parallel reasoning realism is required to make sense of the decision to adopt IRT. This objection makes several presuppositions in treating the two cases analogously, and these presuppositions must be evaluated before the objection's cogency can be assessed. First CTT must not make reference to latent traits as causing item responses or total test score. Second, there must be no *a priori* reason to prefer IRT to CTT. Third, for a given pattern of item responses, it must be that there is an empirically adequate IRT model with $\theta = g$ if and only if there is an empirically adequate CTT model whose true score is identical to general mental ability. The idea behind this presupposition is this: if IRT is the only game in town, there can hardly be a selection to be justified—we lose the *explanandum*. Additionally, if the decision is to reveal essential realist commitments, model selection cannot be confounded by virtues such as the model's ability to fit the data or predictive success that do not require realism. Each of these qualities were exemplified by the case of preferring a reflective model over a formative model.

The first assumption is met if we interpret 'θ' as referring to a latent trait, which seems reasonable so long as we leave open the possibility that 'θ' may denote a cluster of causal confounds and not necessarily some unitary trait. CTT does not make reference to latent traits. It is a tautological theory stating that for any test, there is a "true score" (the candidate trait) defined as the expected score on the test. Test scores measure a trait when scores increase monotonically with the trait (Lord & Novick, 1968, p. 20). Any variable increases monotonically with itself. Borsboom makes the point perspicuous for the case of intelligence tests when he writes:

> …if the true score of an IQ-test is considered to be identical with intelligence, the proposition 'IQ scores measure intelligence' is true by definition. This is because the proposition 'IQ-scores measure intelligence' is transformed to 'the expected IQ-scores are monotonically related to the true scores on the IQ-test' which is vacuously true since the true scores are identical to the expected scores (Borsboom, 2005, p. 141).

Therefore, one has to import additional structure to the theory to get traits out of it; i.e., the traits have to be added in. They are not part of the measurement theory. For those who wish to give CTT a realist interpretation, the burden is on them to distinguish which true scores should be treated realistically and which should not, since for every test there will be an associated true score. This is not the case with IRT models where the latent trait is posited as a partial explanation of test performance.

The second requirement above is not met; that is, there are *a priori* reasons to prefer IRT to CTT. Of course, this is what one would expect seeing as IRT was developed to cope with known problems with CTT. The problems with CTT are well documented and I will not rehearse them at length here.[38] Instead I will highlight only a couple. Under CTT, test difficulty is group-dependent; it is measured as the proportion of examinees to have answered the item correctly. That is to say the difficulty is contingent upon the ability of the group (the greater the number of high ability examinees, the fewer who will get items wrong), but then their ability depends on the difficulty in the sense that a measure of ability will be a consequence of performance, which will depend on how challenging the items are. The same is true of discrimination parameters and reliability—they depend on the group of examinees. Another problem with CTT is that since reliability of a test varies with levels of ability, there is a conceptual obstacle to comparing examinees of different levels of ability *even with respect to performance on the same test*. IRT, which applies at the level of items, is much more precise than CTT which applies at the level of tests and its parameters are not group dependent.

[38] See Hambleton *et al.* for an accessible discussion of the problems with CTT.

The third condition is met in the case of parametric IRT models that posit *g* (and so for unidimensional parametric IRT models generally) (Borsboom, 2005; Crocker & Algina, 1986). In fact there is an equation for computing classical true scores from positions on latent variables.

Let's take stock. Neither IRT nor CTT *requires* ontological realism. It seems that the decision to adopt IRT as a measurement framework as opposed to CTT *does not* offer evidence for methodological realism. The rationality of adopting reflective as opposed to formative models of covariation requires a commitment to minimal epistemic realism, and therein one finds grounds for methodological realism; however, the rejection of CTT in favor of IRT does not require a commitment to minimal epistemic realism to be rational. There are reasons to prefer IRT other than a commitment to the referential success of 'θ' or the possibility of accruing evidence in favor such a claim, and thus ascriptions of epistemic realism are yet underdetermined.

2.23 When Realism *Is* Required by IRT and, in fact, Latent Variable Theory Generally: Estimating Ability

Now I will further complicate things (or clear them up), for I will argue that certain practices within the context of IRT and latent variable theory require at least a tacit commitment to epistemic realism if not full blown ontological realism. Again, since I am investigating the presuppositions of psychometrics and not their justification, I'll leave to the side whether realism is justified.

In section 2.21 I argued that there is a use of IRT models that does not require a commitment to either ontological or epistemic realism, namely the prediction of performance given a level of ability, hence methodological realism finds no vindication in those cases. Metaphysically reticent psychometricians, psychologists, and test-developers may be relieved with these results; after all, agnosticism about the existence of latent traits seems to be the dominant position among those in the relevant sciences. Empiricists, too, will find these results congenial (at the least those that pertain to IRT). Of course, they must still contend with the

results of 2.1. They will be disappointed when they find that IRT in practice, and measurement in psychometrics generally insofar as it appeals to latent variables, cannot escape realist commitments.

Given a function describing an ICC, we can estimate performance on items, given a level of ability and a specification of the parameters of the model, e.g., difficulty. However, it would indeed be strange if tests were administered with only the purpose of determining (probable) performance on test items. The purpose of ability tests is to measure ability. After all, what is the value of saying that John will probably get item 1 correct on test X given that he has ability level A as measured by some other instrument? This seems to get the testing enterprise backward. I don't deny that there is a legitimate place for such a practice in test development; for example, if one's aim were to confirm a model, it would be valuable to see if the model made the correct predictions for performance. But at the end of the day, once an item analysis for the test has been conducted, the model is confirmed, and the appropriate validation studies conducted, the test is then implemented as an instrument for assessing ability in the population at large. Many assumptions go into assigning scores to individuals with respect to some latent factor (e.g., that scores are normally distributed in the population) but I will not be concerned with their legitimacy here, since such a discussion would bear more on the justification of realism than on the present concern. If we grant those assumptions (or turn a blind eye to them) and just focus on what it means to estimate an individual's position on a latent variable, we find that realist assumptions underlie practice.

There are several procedures for estimating someone's *g*-score, none of them trivial. *Maximum likelihood estimates* of θ introduce an operational latent variable θ', which acts as a proxy for θ. Simply put, θ' is an estimate of the value of θ that would render a pattern of item responses most likely. Calculating maximum likelihood estimates is complicated and not always possible. Bayesian estimation procedures may also be used, especially when a maximum

likelihood estimate cannot be calculated. One may also search for an operational latent variable that renders the factor model locally independent, e.g., the sum score; this proxy variable is called a *sufficient statistic* for the latent variable (Bartholomew, 2004). The idea is that once you have identified some function of the $\mathbf{X}_i$, i.e., the manifest test variables, such that it screens them off from one another, you have captured all the information contained in *g*. The sufficient statistic can then be used to place examinees on an ordinal scale.[39] However, the particulars of the various approaches to estimating position on some latent variable need not concern us. It is what they have in common that is relevant to the current discussion: they all carry with them the presumption of ontological realism in addition to a commitment to minimal epistemic realism.

Suppose we administer a test battery, the WISC-IV for example, to two examinees, Nino and Eka. Also, suppose that we have some sufficient statistic θ' for *g*. Eka scores significantly higher than Nino, i.e., she has a higher value for θ' than Nino. We infer from this fact that Eka occupies a higher position on *g* than Nino. We then infer, based on their estimated relative positions on *g*, that Eka is more intelligent than Nino. This latter inference will be trivial if individual ability is defined as position on *g*, or if relative ability is defined as relative position on *g*. For the purposes of illustration, here I will not be concerned with the legitimacy of the latter inference except to point out that estimating ability in this way requires local homogeneity, the claim that the population-level structure of individual differences is isomorphic to the structure of intraindividual differences, since the sufficient statistic will be discovered with population-level data.[40] We then import the sufficient statistic to the assessment of individuals. Whether the assumption of local homogeneity is justified will be discussed in later chapters. And in the next

[39] Ordinal scales give only relative measures. They have neither a non-arbitrary zero point, nor do they permit the interpretation of distances between magnitudes. Ratio scales are not deficient in these respects.

[40] Interpreted causally, the local homogeneity assumption claims that the population-level measurement model represents not only measurement at the population level, but also measurement at the level of individuals. In other words, whatever causes differences in test scores at the interindividual level is also responsible for differences in test scores at the intraindividual level; therefore, if a test measures general mental ability at the population level, it also measures the same trait in each member of the population and each subpopulation.

section, I will show that the dispute over the justification of the assumption of local homogeneity provides further evidence in favor of methodological realism.

There does not seem to be any antirealist interpretation of the practice of estimating individual ability (or the ability of groups). Of course, if the latent trait does not exist or if belief in its existence is unjustified, then any claims of the form 'S's estimated level of some latent trait θ is X' will be unjustified. But the *practice* of estimating latent ability seems to presuppose that there is some dimension on which we locate individuals and that claims about one's position on the dimension is corrigible: the more tests we give, the more confident we can be that S's "true" position on θ is within a certain interval of values. Estimation presupposes that there is a true value on which we are trying to converge (Borsboom, 2005). If psychometricians did not believe that the attribute in question existed, and if they did not further believe that claims about individual position on the attribute were corrigible, the use of estimation procedures and attempts to reduce sampling error by giving a diverse battery of measures would seem to be a perverse and arbitrary enterprise. Consequently, estimating ability requires both methodological realism and a commitment to the existence of latent traits.

If we attend specifically to the procedure where a sufficient statistic is sought, we find further support for methodological realism. Some statistics will be "more sufficient" than others. The reason why a particular statistic is chosen as the sufficient statistic as opposed to another that does not screen off the correlations between the observed measures as effectively is that the former statistic is alleged to contain more information about the latent trait, thus attributions made on its basis will be more justified than those made on the basis of the latter statistic.

2.3. The Fuss over Local Homogeneity Requires Methodological Realism

If a model of covariation is locally homogeneous, then if it fits a population of examinees, then it also fits each member of that population. For example, if either model in Figure 3.1 is locally homogeneous, then it will not only fit the data we obtain from the population, it will also fit each

individual examinee over repeated measures. At this level of description, local homogeneity is just a syntactic relation between models of covariation. However, the interpretation of local homogeneity is that whatever the test measures in the population is also measured in each test taker and subpopulation of test takers (Ellis & van den Wollenberg, 1993, p. 422; Borsboom, 2005), i.e., the covariation models are interpreted causally as measurement models. In testing and test development, local homogeneity is often assumed without justification. However, the justification for making this assumption has come under fire in the past few decades and at least as early as 1946.[41] In spite of this, intelligence and personality theorists persist in making the assumption. I discuss the limitations that a failure of local homogeneity imposes on a conception of intelligence as general intelligence in later chapters. What I show here is that research on local homogeneity and subsequent developments in psychometric theory intended to deal with problems associated with a lack of local homogeneity seem to be motivated by realist commitments.

The force behind a demonstration that a measurement model is not locally homogeneous is straightforward when we consider its interpretation. If local homogeneity is supposed to show that two tests measure the same attribute, then if the population model is not locally homogeneous, the heterogeneity is interpreted as showing that what the test (or battery of tests) measures in the population is not identical to what is measured in each individual. In the fields of personality and ability assessment, the alleged objects of measurement are latent traits such as extroversion and general intelligence. Therefore, to the extent that a test is taken to measure an attribute in the population (or individual), ontological realism is assumed. Moreover, to the extent that a failure of local homogeneity is supposed to be evidence that what is measured in the population is (or may not be) not what is measured in the individual, epistemic realism is assumed.

[41] Baldwin writes, "There is no assurance, however, that the organization of personality variables within the individual is accurately described by the pattern derived from group studies," (1946, p. 152).

In response to the problem posed by local homogeneity, psychometricians have developed and implemented statistical tools for the study of individuals over time (*cf.* Cervone, 2004; Hamaker, 2004; Molenaar, 1985, 1999; Molenaar, Huizenga, & Nesselroade, 2003). Analysis of intraindividual variability over time is conducted through *time-series analysis*. Hamaker in particular suggests that psychology should shift from its current focus on interindividual variability, studied by way of standard factor analytic procedures, to intraindividual variability, studied by way of dynamic factor models reflecting change at the level of the individual. The thought is that through time-series analysis and its integration with traditional methods, we will gain insight into psychological processes. For example, Hamaker writes,

> If we want to obtain knowledge about intraindividual psychological processes and about the lawfulness that underlie them, we must employ a technique that allows us to make statements about the structure of variability within the individual, rather than about the distribution of variables in the population…If the individual is our unit of analysis, the results [of population-level data] do not inform us about what is going on at the level of the individual, but about the laws operating at the level of the population (Hamaker, 2004, p. 7).

The above quotation is typical of those who advocate a turn from traditional nomothetic studies in which the unit of analysis is the population, to ideographic studies in which the unit of analysis is the individual. However one need not be an advocate of time-series analysis to acknowledge the statistical fact that some models are not locally homogeneous. Time-series analysis and hybrid approaches such as Hamaker's have not enjoyed immense popularity, but it is unclear whether this is for sociological reasons such as the complexity of the techniques and the fact that standard statistical software packages popular in the social and behavior sciences (e.g., SPSS and SAS) do not currently include the requisite statistical techniques, or whether it is because the techniques are inappropriate from a methodological perspective. Nevertheless, there does seem to be a significant contingent within psychometrics advocating an ideographic turn in research and analysis. My claim is that the impetus for this turn is realist in character.

Let us first consider the epistemology of the ideographic turn. Recalling Hamaker's quote above one can see that it wears its epistemic realism on its sleeve, claiming that "if we want to obtain knowledge" of individuals' psychological processes, then we ought to employ techniques appropriate for studying intraindividual change. Nevertheless, we don't have to take Hamaker's (or any other psychometrician's) word for it. After all, the assumption of epistemic realism may turn out to be otiose. But what would motivate the ideographic turn if not a commitment to the possibility of gaining evidence about intraindividual trait structure and how traits (or psychological attributes) change over time? Without this commitment, the ideographic turn seems unmotivated. It is highly implausible that psychometricians and psychologists would campaign for using what are incredibly sophisticated and complex techniques without implicitly acknowledging the possibility of delivering evidence concerning the nature of psychological attributes. And for those who take the turn, it cannot be rationalized in terms of simplicity or pragmatics, for the techniques in question are not part of the psychologist's or psychometrician's standard toolset, nor are they easily implemented.

For the same reason that modeling covariation reflectively does not require a commitment to ontological realism, neither does the ideographic turn. That is, each factor structure at any point in a time-series may be considered a hypothesis subject to confirmation or refutation. Indeed the entire dynamical model of an individual's psychological change may be so considered. It would seem odd if one were propounding such hypotheses without any regard for their truth (or representational adequacy); presumably one formulates hypotheses and conducts empirical tests with the aim of one day accepting some hypothesis as (at least partially) true, but this aim amounts to methodological realism. The impetus for the turn is that time-series analysis offers better and more nuanced hypotheses, ones that are more likely to represent intraindividual psychological processes. Nevertheless, one may never meet the epistemic standards whose fulfillment warrants ontological realism, so ontological realism is not necessary to make sense of the ideographic turn.

3. Psychometrics as Pathological Science: Gould and Michell

In *The Mismeasure of Man* Stephen Jay Gould argues that since the initial development of factor analysis and the "discovery" of *g* by Charles Spearman in 1904 (Spearman, 1904), psychometricians have unreflectively and blatantly reified general intelligence. Psychometricians have responded to Gould by saying that his knowledge of psychometric methods is decades out of date, and they have responded by professing agnosticism about the nature and existence of latent traits, regarding them as hypotheses. If what I have said in the previous sections is true, then there might yet be grounds for Gouldians to object. Gould was content to attack specific psychometricians for their ontological commitments; those to whom he objects most vociferously I exonerated in chapter 1. What I propose here on Gould's behalf is, if cogent, a much more serious objection since it implicates modern psychometrics itself, not simply a handful of its practitioners. Thus I have sought to update the Gouldian objection in response to those who would complain that the target of his polemic is a 30-year old straw man.

A brief note regarding Gouldians and methodological realism: they are not likely to balk at the result that methodological realism finds support in much of psychometric practice. The philosophical significance of this result is that psychometrics cannot be rationally reconstructed without realist commitments. In one sense, this can be seen as the starting point of Gould's objections. However, were I to *defend* epistemic or ontological realism in psychometrics, at that point Gould and I would have to part ways since his attack on the reification of latent factors aims to undermine realism, both ontological and epistemic.

The spirit, if not the letter, of Gould's objections seems correct: the unsubstantiated reification of mathematical entities seems at odds with legitimate scientific practice. More recently, Michell (1999, 2000, forthcoming) has charged that IRT theorists are particularly guilty of this transgression in that they assume that latent traits exist, have a quantitative structure, and that those theorists are unresponsive to criticisms of this assumption. For this reason Michell

considers psychometrics to be an instance of "pathological science." For a science to be pathological two conditions must be met. First, a hypothesis must be accepted as true without evidence. Second, there must be no acknowledgment that there is such an assumption or it must be disguised (Michell, forthcoming). Here I am assuming that the unwarranted reification of latent statistical factors in the form of psychological traits or attributes, especially when insulated from criticism in the form of recalcitrant evidence, is an instance of unresponsiveness to criticism.

Gould and Michell's general and very sensible idea that unresponsiveness to criticism represents a breakdown of scientific objectivity finds considerable support in recent work in the philosophy of science (*cf*. Fagan, 2007; Kitcher 2001; Lloyd, 2005; Longino, 1990, 2002). So the question is this: does Gould's and Michell's criticism have a target in IRT or any of the examples discussed above? It seems not. Ontological commitment in absence of epistemic realism would be sufficient for unresponsiveness to criticism, since it renders ontological commitments incorrigible. However, ontological realism in conjunction with epistemic realism seems to be perfectly acceptable, for it entails a willingness to allow empirical commitments to answer to evidence. Epistemic realism alone also seems entirely benign, for at worst, i.e., if epistemic antirealism is true, epistemic realism would be tantamount to an overly optimistic epistemic aspiration, but it would still place empirical evidence as a key arbiter in empirical disputes. I take it that if epistemic realism is unscientific or pathological, then much of what is regarded as exemplary science is rendered unscientific, and that is unacceptable.

At this point, one may object that I have attended to the wrong aspect of Michell's objection, that what he is objecting to is not the ontological commitment to latent traits as such.[42] Rather, it is that psychometricians assume that latent traits, if they exist, have a quantitative structure. It is the quantity assumption that is under attack, not the postulation of the entity itself (insofar as we can separate the postulation of the entity and the properties we ascribe to it). Further, it may be objected that I have read Michell too strongly in interpreting him as indicting

[42] The following response to Michell appears in Hood, S. Brian (forthcoming) only slightly altered in form.

psychometrics as a discipline. The actual targets of Michell's criticism are individual psychometricians. So, his objection does not claim that psychometrics is somehow intrinsically pathological. Michell is actually making a sociological claim about psychometric practice as done by individual psychometricians. The pathological nature of psychometrics stems the way psychometrics is practiced. I will consider these objections in turn.

There seems to be at least one counterexample to Michell's conception of pathological science. The background assumptions of a scientific community (such as the quantity assumption) may not be transparent to its members. Since background assumptions figure in determining (among other things) what counts as evidence and the relation between evidence and theory, their role is significant. But if background assumptions are not transparent to their adherents then we have satisfied the criteria for pathology. Unfortunately, Michell indicts much of the rest of science, for it is often the case that a community's background assumptions are not transparent to its members (*cf.* Longino 1990).

I agree with Michell that once the background (quantity) assumption is made explicit and is challenged by critics, the *relevant* community has a responsibility to respond to that criticism. However, one place where I differ with Michell is in what constitutes an adequate response. For Michell, acknowledging the criticism is adequate. Michell claims that "Any group of psychometricians would be exempt were they to admit that they *assume* the empirical hypothesis that psychological attributes are quantitative… [by] saying something like, 'at present we do not know whether this hypothesis is true, but we will assume it recognizing that at some point in the future someone needs to investigate it[,'" (Michell, forthcoming). So, according to Michell, if every paper aimed at measuring psychological attributes simply tacked on the nested quotation above, psychometrics would cease to be pathological. First, it is implausible that psychometricians aware of Michell's criticism and the relevant background assumption would deny that "at some point in the future someone needs to investigate" whether psychological attributes are really quantitative. Second, if merely admitting to making certain assumptions yet

to be confirmed is sufficient to avoid charges of being pathological, then Michell's bark seems much worse than his bite. However, again, demanding recognition of one's own background assumptions is perhaps too much to ask (though responding to criticisms once they are made explicit is not). Michell sometimes claims much more, namely that it is because the tests for the quantitative structure of attributes *have not been done*, that psychometrics is pathological. But note that this is a stronger demand than the initial one, *viz.* that *merely acknowledging* the assumption and stating (in print apparently) that it should be tested is sufficient to dispatch charges of doing pathological science.

Briefly I would like to remark on the semantic and consequent normative import of the notion of "pathological science". It is unclear what lies in the domain of the predicate 'is pathological'. The title of Michell's latest contribution, "Is Psychometrics Pathological Science?" suggests that it is all of psychometrics, though at times he writes as though it is particular psychometricians who are doing (or not doing) psychometrics pathologically. I take it that Michell's main claim is not as strong as the title of this piece suggests. To indict all of psychometrics as pathological would be a bit hasty. There is no intrinsic feature of psychometrics *as a discipline* that justifies a wholesale charge of being pathological. This would do an injustice to those psychometricians whose work is not pathological according to Michell's criteria. Michell gives examples of psychometricians who acknowledge the quantity, so we can safely say that that their research is in the clear. Thus, I submit that Michell's target is individual psychometricians. This interpretation is less problematic than that which would have him indicting all of psychometrics; it also accords well with his claims in Michell (2004).

Michell exposes a strong empirical assumption underlying much of psychometric research, and with this background assumption made explicit it is incumbent upon someone to test the assumption. Here I agree with Michell, but it isn't clear who should do this work. Perhaps this is why it has not been done. But if it is unclear whose responsibility it is to do the required

work, i.e., what the relevant community is, then it is unclear who is to blame for not doing it, and it is thus unclear who, exactly, is doing pathological science.

In none of the examples considered above has a case of ontological realism divorced from epistemic realism been encountered. In the case of model selection, I argue that we find evidence for methodological realism, but not necessarily a presupposition of ontological realism; in the case of using IRT models to predict item or test performance, no realist assumptions are required; in the case of adopting IRT in favor of CTT I argue that the case underdetermines methodological realism and that the adoption of IRT need not indicate an assumption of ontological realism; in the case of using IRT models to estimate ability I argue that there is an implicit commitment to ontological realism, but that it is accompanied by a commitment to epistemic realism; in the assessment of the problem posed by non-locally homogeneous measurement model I likewise argue that the commitment to ontological realism accompanies a commitment to epistemic realism; and in the case of the ideographic turn I argue that the development of statistical tools for the analysis of intraindividual variability evidences a commitment to epistemic realism, but not necessarily ontological realism. Psychometricians will have various philosophical commitments, but there is nothing in the contemporary methods discussed here that requires a pathological ontological commitment.

For those epistemic (or semantic) antirealists who remain unconvinced by my argument that certain methodological practices require (or are at least most plausibly interpreted as betraying) realist presuppositions, I should add that antirealism provides no safe haven. The most obvious targets for Michell's critique are ontological realists, for they, if anyone, endorse the quantity hypothesis. My arguments for the necessity of realist presuppositions can be read as an attempt to locate Michell's most obvious opponent, but realists alone do not bear the burden. Antirealists may respond to Michell's critique claiming that they grant no such endorsement; they will seek exemption claiming that they are uncommitted to the existence of quantitative latent traits. However, Michell's critique stands, if only at the level of methodology. The target of the

critique can be shifted from the psychometrician's ontological commitments to methodological practice to undermine the antirealist response. For example, suppose that I am wrong, and the estimation of ability in IRT can be plausibly interpreted in an antirealist light. I suppose the antirealist story would go something like this: the practice does not indicate an underlying commitment to the existence of the latent ability; what it indicates is a preference to model data *as if* there were such an ability. I think that most psychometricians would protest that this grossly distorts how they think of measurement models, and I argue that regardless of one's professed commitments, this practice is realist. Nevertheless it is conceivable that a philosopher occupies such a position. Michell could so tailor his objection: proceeding *as if* the traits were quantitative is no better than assuming that they are. From a methodological standpoint there is no difference. The quantity hypothesis does the same work in the models even if it does not require a realist interpretation; it, too, requires testing just as any other auxiliary hypothesis would. But going along with the "*as if*" antirealist even this far might be too charitable. It is unclear how the "*as if* antirealist" could make sense of testing the quantity hypothesis without being committed to at least epistemic realism. Nevertheless, to the extent that "*as if*" antirealism is a coherent position in this context, Michell's critique stands.

I'd like to conclude this section with a brief aside. It may be argued that dogmatic commitment to the existence of theoretical posits may not be regarded as a failure of objectivity or as pathological. For example, Imre Lakatos (1970) argues that it is such a commitment that characterizes a scientific research program and it is necessary for scientific progress. All scientific research programs, including those deemed 'progressive' on Lakatos' account, have a theoretical "hard core" which is insulated from recalcitrant evidence and refutation by a "protective belt" of auxiliary hypotheses. If one accepts Lakatos' methodology of scientific research programs, then he is unlikely to object to dogmatic ontological commitments, consigning them to the hard core of the program.

4. Conclusion

In an attempt to locate a suitable and clear target for Michell's critique of psychometrics as an instance of pathological science and to argue that methodological practice is, at least in some cases, essentially realist, I have argued for methodological realism in psychometrics—that certain practices and methodological shifts in psychometrics seem to depend on realist commitments for their rationality. These commitments stand apart from the philosophical positions of particular psychometricians and even may be inconsistent with their professed commitments since background assumptions are not always clearly manifest, not even to those who harbor them. In some instances practice seems to indicate realist commitments where, upon closer inspection, there are none. In other instances close examination reveals a tacit commitment to ontological realism, epistemic realism, or both, though I argue that in none of the examples considered here do we find a commitment to ontological realism without a corresponding commitment to epistemic realism. Ontological realism unwarranted by evidence would provide a case for judging psychometrics as a "pathological science," and it was this pathology that so concerned Stephen Jay Gould. I argue on the basis of several considerations, including the problematic nature of the notion itself, that the charge that psychometrics is pathological finds no ground in the examples provided. Further, I argue that the assumption of epistemic realism is innocuous and represents a willingness to listen to evidence, i.e., a form of Helen Longino's "responsiveness to criticism" requirement for scientific objectivity. An additional consequence of my arguments here is that there are certain features of psychometric theorizing that cannot be explained on an empiricist account, and if attempting to make sense of science is one of the tasks of the philosophy of science, the empiricist is bound to fail in his attempts in the case of psychometrics. Realism affords the epistemic and metaphysical richness needed to make sense of psychometric practice and developments. Contemporary psychometrics is far from being the atheoretical offspring of operationalism that it is often alleged to be. Nevertheless, should one insist that psychometrics *can* be understood from an empiricist standpoint, one still does not escape Michell's critique

simply in virtue of antirealist epistemic and metaphysical reservations; parsimony is no cure for pathology.

CHAPTER FOUR

A Comparison of the Bifactor and Higher Order Factor Models of Intelligence: Philosophical and Psychometric Considerations

What you have in your head, put it down on paper. The head is a fragile vessel.

Dmitri Shostakovich

1. Introduction

In this chapter, I argue that the question of whether one is a realist bears on model selection in psychometrics. To illustrate this point, I consider two different confirmatory models of intelligence data: the bifactor model and the higher order factor model. Briefly, my goals are twofold: to answer to psychometricians who argue that bifactor models are preferable to higher order factor models, and to show that the question of realism matters in adjudicating between these two models. I do not argue for the truth of realism or even its plausibility. Rather, I wish to show that in psychometrics, for the purposes of statistical modeling, it matters if one is a realist.

Bifactor models (Holzinger and Swineford, 1937), also referred to as "general-specific models" or "nested factor models" (Gustafsson and Balke, 1993), and higher order factor models offer competing psychometric representations of the covariance structure of intelligence test scores. Specifically they differ in the manner they accommodate the general intelligence factor (Jensen, 1998). Bifactor models posit the general factor and group factors as orthogonal first order common factors. In contrast, higher order factor models posit the general factor as a second order factor[43], which accounts for, or explains, the covariance among the first order factors. Recently Gignac (2005a, 2005b, 2006) and Chen *et al*. (2006) have argued on the grounds of psychometric and statistical considerations such as goodness of fit, utility, and conceptual tractability, that bifactor models are preferable to higher order factor models; Gignac makes his arguments specifically in the context of intelligence research. A notable proponent of the higher order factor model is Jensen, who states that "[a]mong the various methods of factor analysis that do not…preclude the appearance of *g* [i.e., the general intelligence factor]…a hierarchical model is generally the most satisfactory, both theoretically and statistically," (Jensen 1998, p. 73).

Both the higher order factor model and the bifactor model are applied in the intelligence literature and presented as ways of modeling the relation between general intelligence, on the one hand, and intelligence test data, on the other. The precise representation of such relations is a matter of substantive importance, because the theoretical ideas laid down in the theory of general intelligence can be tested only through statistical models; therefore it is crucial that, in testing one's theory, the statistical model that one uses is actually a representation of that particular theory, and not of some other theory. Now, it appears as if scholars who have considered the relative merits of the bifactor and higher order factor models have done so mainly on the basis of pragmatic considerations, apparently assuming that the two models gave equally adequate representations of general intelligence. Few have noted that there is a far more important

[43] To ease presentation, discussion is limited to first and second order factor in the higher order factor model. It should be noted that the general intelligence factor may be specified as a third order factor (e.g., see Johnson, *et al.*, 2003; Carroll, 1994).

consequence of selecting a particular model; namely, the choice of model determines *what particular theory one is testing*. The reason is that, as will be shown in the present paper, the higher order and bifactor models do *not* offer exchangeable statistical representations of general intelligence; that is, the higher order model's *g* is not the bifactor model's *g*. The same holds for the group factors in the higher order model and the residual group factors in the bifactor model; the former may be interpreted as well-defined, domain specific cognitive abilities, however the latter may not.

The question then naturally arises which model's *g* is the right *g*, in the sense of being an adequate statistical representation of the substantive theory of general intelligence as articulated by Jensen (1998). I will argue that this honor falls to the higher order factor model. The focus on Jensen's theory is motivated by the fact that he is among the most prolific and influential advocates of the *g* factor and the theory of general intelligence. The immediately relevant follow up question, of course, is what theory the bifactor model is representing if it is not Jensen's theory of general intelligence. I do not have the answer to this question, but I do present a number of statistical and theoretical implications that follow from assuming the truth of a bifactor model in order to carve out the logical space in which the relevant theory of intelligence should situated. However, the relevant implications appear to be implausible to the extent that the chances of finding a reasonable concomitant theory for the bifactor model may be considered surprisingly slight. For this reason in conjunction with certain pragmatic virtues of the bifactor model, I suggest that it may be preferable to interpret the model in a purely instrumentalist sense.

The structure of this chapter is as follows. First, I will present three confirmatory factor models of intelligence data: the oblique first order factor model, the higher order factor model, and the bifactor model. Second, I will note the alleged pragmatic virtues of the bifactor model including goodness of fit, the ease of detection of redundancy in terms of the number of components in residual group factors, the accommodation of external regressors of the common factors in the model, and testing for measurement invariance. I argue that except perhaps with

regard to goodness of fit, there are reasons to prefer the higher order factor model. The arguments assume a realist stance toward the measurement model, though they do not all rely on the presumption of realism for their cogency. In adopting a realist stance, I treat the measurement model as representing an actual causal relationship between well-defined, domain specific cognitive abilities and indicators such as item or test performance. Third, I will analyze theoretical implications of the bifactor model, arguing that they are incompatible with at least one mainstream theory of general intelligence, and that the bifactor model presently lacks a concomitant psychological theory that would make sense of some of its specific consequences, e.g., that it makes the construction of unidimensional and measurement invariant tests of cognitive ability impossible. I then point out significant epistemological differences between first order and second order latent variables and marshal these differences to defend the claim that the general factor in the bifactor model is distinct from the general factor in the higher order factor model. Finally, I consider further epistemic considerations relevant to the higher order factor model.

2. Confirmatory Models of Intelligence Data

In this section, I present three confirmatory factor models of intelligence tests: the oblique first order factor model, the higher order factor model, and the bifactor model (Gustafsson & Balke, 1993; Bollen, 1989; Rindskopf & Rose, 1988). I include the oblique first order factor model as it provides a conceptually and statistically important baseline model. In the interest of being thorough, the models are presented in the matrix notation (based in part on the LISREL (Linear Structural Relationships) model notation). LISREL is the standard statistical software package used in confirmatory analyses. The models are then given path diagrammatic representations. I also present recent arguments for accepting the bifactor model as a model of intelligence.

2.1 The Oblique First Order Factor Model

Let $\mathbf{y}$ denote the p×1 random (i.e., arbitrary) vector of observed intelligence test scores (omitting subject subscripts). Let $\boldsymbol{\eta}$ denote the q×1 random vector of common factor scores, upon which $\mathbf{y}$ is regressed:

$$(1) \qquad \mathbf{y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\,\boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\tau}$ is a p×1 vector of intercepts, $\boldsymbol{\Lambda}$ is a p×q matrix of factor loadings, and $\boldsymbol{\varepsilon}$ is a p×1 random vector of residuals. Assuming that the mean of $\boldsymbol{\varepsilon}$ is zero, and $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are uncorrelated, the expected mean vector, $\boldsymbol{\mu}_y$, and covariance matrix, $\boldsymbol{\Sigma}_y$, are (e.g., Lawley & Maxwell, 1991; Bollen, 1989):

$$(2) \qquad \boldsymbol{\mu}_y = \boldsymbol{\tau} + \boldsymbol{\Lambda}\,\boldsymbol{\mu}_\eta,$$

$$(3) \qquad \boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}\,\boldsymbol{\Sigma}_\eta\,\boldsymbol{\Lambda}^t + \boldsymbol{\Theta},$$

where $\boldsymbol{\mu}_\eta$ is the q×1 mean vector of the common factors, $\boldsymbol{\Theta}$ is the p×p positive definite (p.d.) covariance matrix of the residuals $\boldsymbol{\varepsilon}$, and $\boldsymbol{\Sigma}_\eta$ is the q×q p.d. covariance matrix of the common factors $\boldsymbol{\eta}$. We assume that $\boldsymbol{\Theta}$ is diagonal (in the sense that all the off-diagonal entries in the matrix are zero), and the matrix $\boldsymbol{\Lambda}$ displays simple structure, i.e., if a variable loads on one first order factor, it has a loading of zero on the other first order factors. These assumptions, which are made in order to interpret that data, may be considered idealizations, which may be relaxed to various degrees. The path diagrammatic representation of the oblique first order factor model, or "group-factor model" (Rindskopf & Rose, 1988) is shown in Figure 4.1.

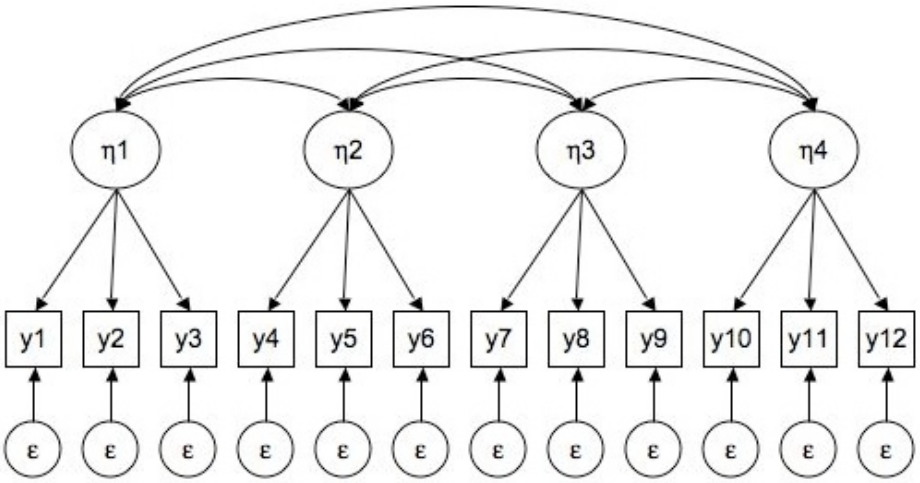

Figure 4.1: Oblique first factor model with four group factors η, twelve indicators **y** and their associated residual terms ε.

2.2 The Higher-Order Factor Model

Noting that common factors are invariably positively correlated in analyses of intelligence test data, one may fit what I will call the higher order factor model (henceforth "the HF model"; Jensen, 1998). This model may be viewed as an elaboration or extension of the oblique first order factor model. The presentation is limited to a second order model, in which the covariance structure of Ψ, is the q×q diagonal covariance matrix of the residuals, is modeled by introducing a single second order common factor. This is achieved by introducing the regression model:

$$\eta = \Gamma \xi + \zeta, \quad (4)$$

where ξ is the second order factor score (here this will represent general intelligence or *g*), Γ is the q×1 matrix of regression coefficients, and ζ is the q×1 random vector of residuals. Assuming

ξ and ζ are uncorrelated, we obtain the following covariance matrix of the first order common factors, $\boldsymbol{\Sigma}_{\eta}$:

$$(5) \qquad \boldsymbol{\Sigma}_{\eta} = \boldsymbol{\Gamma}\,\boldsymbol{\Phi}\,\boldsymbol{\Gamma}^{t} + \boldsymbol{\Psi},$$

where $\boldsymbol{\Phi}$ is the 1×1 covariance matrix of ξ, and $\boldsymbol{\Psi}$ is the q×q diagonal covariance matrix of the residuals. The mean vector, $\boldsymbol{\mu}_{\eta}$, is

$$(6) \qquad \boldsymbol{\mu}_{\eta} = \boldsymbol{\Gamma}\,\boldsymbol{\kappa} + \boldsymbol{\alpha},$$

where $\boldsymbol{\kappa}$ is the mean of ξ and $\boldsymbol{\alpha}$ is a q×1 vector containing the means of ζ. Combining eqs. 2, 3 and 5, 6, we have:

$$(7) \qquad \boldsymbol{\mu}_{y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\,[\boldsymbol{\Gamma}\,\boldsymbol{\kappa} + \boldsymbol{\alpha}],$$

$$(8) \qquad \boldsymbol{\Sigma}_{y} = \boldsymbol{\Lambda}\,[\boldsymbol{\Gamma}\,\boldsymbol{\Phi}\,\boldsymbol{\Gamma}^{t} + \boldsymbol{\Psi}]\;\boldsymbol{\Lambda}^{t} + \boldsymbol{\Theta}.$$

The path diagrammatic representation is depicted in Figure 4.2. An alternative representation (Schmid and Leiman, 1957) of the covariance matrix is as follows. Let $\boldsymbol{\Omega}$ represent the (q+1 × q+1) diagonal matrix

$$\boldsymbol{\Omega} = \begin{matrix} \boldsymbol{\Phi} & \mathbf{O} \\ \mathbf{O}^{t} & \boldsymbol{\Psi}, \end{matrix}$$

where **O** is a 1×q zero matrix. Letting [**ΛΓ** |**Λ**] denote the p×(q+1) partitioned factor loading matrix, we have

$$\Sigma_y = [\Lambda\Gamma\,|\Lambda]\,\Omega\,[\Lambda\Gamma\,|\Lambda]^t + \Theta. \quad (9)$$

This is called the "orthogonalized higher order factor model" (Jensen, 1998), as it presents the HF model as an orthogonal first order factor model. The transformation from eq. 8 to eq. 9 was also derived by Schmid and Leiman (1957; see also Wherry, 1959) in the exploratory model. The path diagrammatic representation is given in Figure 4.2.

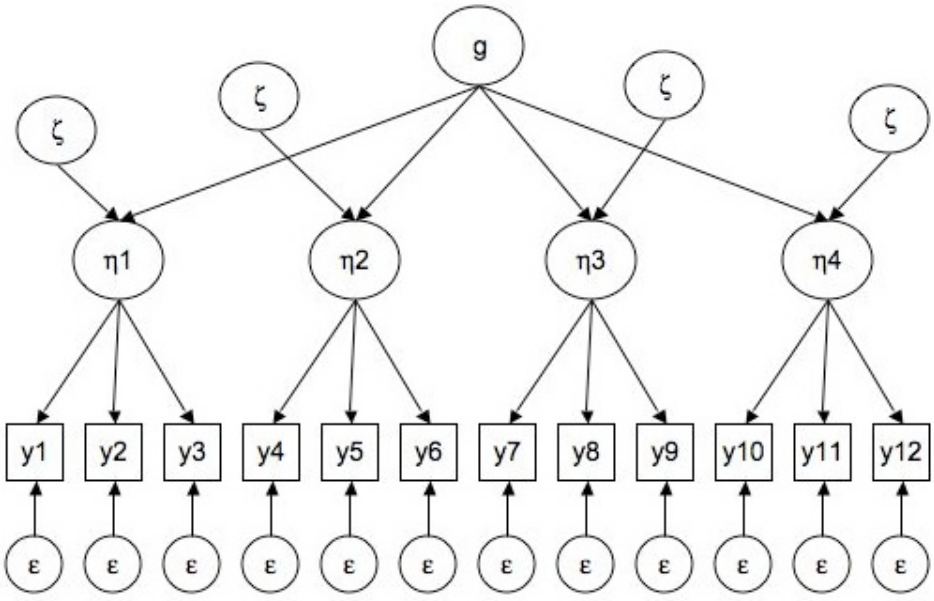

Figure 4.2: Higher order factor model of general intelligence. '*g*' denotes the general intelligence factor and '**ζ**' denotes residuals associated with first order factors **η**. As in Figure 4.1, there are twelve indicators **y** and associated residual terms **ε**.

2.3 The Bifactor Model

The bifactor (BF) model represents a q+1 orthogonal common factor model. The q+1 common factors include one general common factor, upon which all intelligence test scores load, and q

group factors, representing the residual group factors, $\boldsymbol{\zeta}^*$.[44] Let the matrix $\boldsymbol{\Lambda}_1$ denote the p×1 matrix containing the factor loadings of the indicators on the general common factor, and $\boldsymbol{\Lambda}_q$ denote p×q matrix of the loadings of the indicators on the residual group factors. We can then present the model as

$$\boldsymbol{\Sigma}_y = [\boldsymbol{\Lambda}_1|\boldsymbol{\Lambda}_q]\, \boldsymbol{\Omega}^* \, [\boldsymbol{\Lambda}_1|\boldsymbol{\Lambda}_q]^t + \boldsymbol{\Theta}, \quad (10)$$

where $\boldsymbol{\Omega}^*$ is a (q+1 × q+1) diagonal covariance containing the variances of the general common factor and the variances of the residuals in $\boldsymbol{\zeta}^*$, and the p×q factor loading matrix $\boldsymbol{\Lambda}_q$ is again assumed to have simple structure. The BF model is depicted in Figure 4.3.

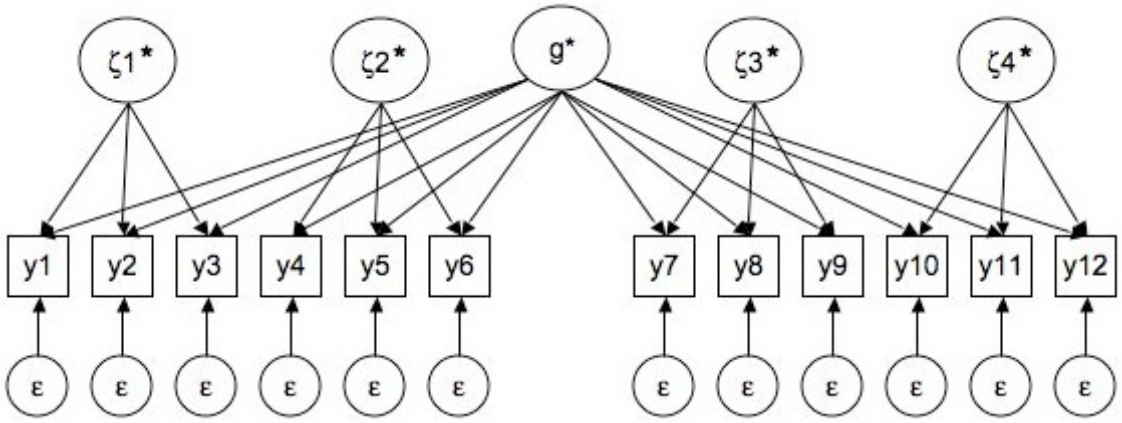

Figure 4.3: Bifactor model of general intelligence. 'g^*' denotes the general intelligence factor and '$\boldsymbol{\zeta}^*$' denotes residual group factors. As in Figures 4.1 and 4.2, there are twelve indicators **y** and associated residual terms $\boldsymbol{\varepsilon}$.

In fitting these models, scaling constraints are required to fix the scale of the latent factors. These constraints may take various forms (e.g., Bollen, 1989). As the exact choice is

[44] I use the term "residual group factor" to stress that the common factor $\boldsymbol{\zeta}^*$ does not represent a unitary cognitive ability (say "Verbal Comprehension"), which is defined by the indicators. Rather it represents residual common factor after accounting for the direct effects of g^* on the indicators. It is an open question whether this residual group factor can be interpreted in terms of a cognitive ability per se. I return to this issue later.

immaterial to the present concerns, I simply assume that appropriate scaling constraints are imposed. In practice, distributional assumptions are made concerning the observed variables **y**. Often the data are assumed to be normally distributed, and normal theory maximum likelihood estimation (ML) is used to fit the models. For present purposes, distribution of the observed variables is immaterial.

Before evaluating these models, note the following. The HF model (eq. 8 or 9) is nested under the oblique first order factor model (eq. 3), i.e., eq. 3 reduces to eq. 8 by imposing eq. 6. Furthermore, the HF model is nested under the BF model, i.e., eq. 10 reduces to eq. 9 by imposing $\mathbf{\Lambda}_1 = \mathbf{\Lambda\Gamma}$ (Yung *et al.*, 1999). Given appropriate distributional assumptions, these nestings permit the application of the standard likelihood ratio test (Azzelini, 1996; Bollen, 1989). In the absence of any additional constraints, the likelihood ratio test of the HF model vs. the oblique common factor model requires at the presence of at least 4 first common factors. Furthermore, it is clear that the BF model and the HF model both include a general common factor (general intelligence). However, it should be noted that the general common factors in the BF model are not equivalent to the common factors in the HF model. To emphasize the distinction between these general factors, I designate the general common factor '*g*' in the HF model, and 'g^*' in the BF model. Similarly, I distinguish between the residual $\boldsymbol{\zeta}$ in the HF model, and $\boldsymbol{\zeta}^*$ in the BF model. Note that *g* and g^* are identical (i.e., represent the same latent variable) and $\boldsymbol{\zeta}$ and $\boldsymbol{\zeta}^*$ are identical (represent the same residuals), if and only if the constraint $\mathbf{\Lambda}_1 = \mathbf{\Lambda\Gamma}$ holds. In all other cases, the general factors in these models are distinct. This motivates a theoretical analysis of the substantive hypotheses that are involved when shifting from one model to the other, an issue that we discuss later in this paper. However, first I provide an overview of methodological, statistical, and pragmatic arguments that have been proposed for and against the models considered.

3. Pragmatic Virtues

In defense of the BF model, Gignac (2005a, 2005b, 2006), Chen *et al.* (2006), and Carroll (1997) have argued that the BF model provides a better fit to the data relative to the HF model, that it is a better model for exploring factor structure due to its greater flexibility, and that in using a BF model it is easier to detect "pure" indicators of general intelligence. Gignac (2005a) fitted the BF model to data from the WAIS-R normative samples and, in comparing the relative goodness of fit of the BF and HF models, found that the BF model fitted best. Specifically, the BF model, comprising three first order factors and a general factor, was associated with superior goodness of fit measures on the standard measures (i.e., Comparative Fit Index, Root Mean Square Error of Approximation, Standardized Root Mean Residual, and Tucker-Lewis Index; see Schermelleh-Engel, Moosbrugger, & Müller, 2003).

Chen *et al.* (2006) noted several other advantages of using the BF model. These relate to the detection of redundancy in terms of the number of components in ζ^*, the accommodation of external regressors of the common factors in the model, and tests of measurement invariance, which I address in the following section. With respect to the detection of redundancy in terms of the number of components in ζ^* it may be desirable to establish that the variance of a given component of ζ^* approaches zero, i.e., that the general common factor g^* completely accounts for covariances among a set of indicators. Chen *et al.* (2006) argued this is evident in the BF model as it results in low and insignificant loadings of the indicators on the component of ζ^*. In the HF model, this is supposed to be more difficult as the loadings are specified on the common factor η rather than on the residual ζ (see eqs. 1 and 4; compare Figures 4.2 and 4.3). Whether, in fact, a test of ζ is more difficult than a test of ζ^* is disputable on the grounds that in the HF model a likelihood ratio test of ζ involves a univariate test, whereas the analogous test in the BF model is a multivariate test equal in number to the number of indicators as it concerns all the factor

loadings of all the indicators of ζ^*. Whether in fact the hypothesis that a component of ζ^* can be dropped is easier to test than the hypothesis that a component of ζ can be dropped is debatable. The likelihood ratio test in the latter case concerns a single parameter, i.e., σ_ζ^2, and the null distribution of the test statistic (i.e., under $\sigma_\zeta^2 = 0$) is non-standard, but relatively simple, viz, a mixture of a $\chi^2(0)$ and a $\chi^2(1)$ with equal mixing proportions, i.e., .5: .5 (Stoel, *et al*., 2006). Given, m indicators of the component of ζ^*, the likelihood ratio test is necessarily an omnibus test, as it concerns as m factor loadings in the matrix Λ_q (see eq. 10). The null distribution of the test statistic in this case is non-standard, and complex, i.e., a mixture of χ^2 distributions, with mixing proportions that depend on the information matrix (*cf*. Stoel, R. D., *et al*., 2006)

The second advantage concerns the utility of the BF model in determining the regression relationships between the group factors and external criteria independently of the general factor. While this can also be done in the HF model, the procedure is claimed to be more straightforward in the context of the BF model, for it does not require the use of nonstandard structural equation models. That is, in the BF model, the residual group factors ζ^* are accommodated as orthogonal first order factors, which may be employed readily as predictors of dependents on external criteria; the BF model does not require employing statistical techniques that are not part of standard statistical software packages and that might not be familiar to other researchers. Note that this argument would speak only in favor of *using* one model as opposed to the other. It is not a shortcoming *of the model* that it may require the use of nonstandard statistical tools; this is merely a sociological fact about the practices of statistical modelers. It does not speak to the *truth* of the model that it is easier to use. Such considerations might recommend that the less complicated model be used when the two models converge with respect to their predictions or when the simpler model makes predictions that are close approximations of the predictions derivable from the more complicated model. Such is commonplace in the physical sciences where

Newtonian models of phenomena are employed when their predictions approximate those of relativistic models. For example, engineers do not use relativistic physics in constructing bridges or airplanes; the added complexity does not yield appreciable differences in physical measurement in the context of building bridges or airplanes. Newtonian models, though demonstrably false, are useful surrogates for relativistic models under certain physical constraints. However, the relative simplicity of Newtonian models is no more reason to think that they are true than is the complexity of relativistic models for thinking that they are false. Physicists and engineers use Newtonian models regularly, but no one suggests that they are true simply because the alternative is so much more difficult to use. The same goes for the case of the BF model versus the HF model. Even if the BF model is easier to use than the HF model, this does not constitute an argument that the former is more likely to be true than the latter, even if results do happen to converge in some (limit) circumstances. But as we will see, the two statistical models are not equivalent with respect to their psychological significance, not even the limiting case where they are statistically equivalent.

To those who believe that *all* models are false by design, the analogy may not be compelling; however, in response I would claim that while all models may be false, strictly speaking, some are closer approximations to the truth than others. Furthermore, the arguments in this chapter take realism regarding the measurement model as a starting point. Needless to say, unless utility were the sole consideration, i.e., if one is an instrumentalist and does not count truth as a theoretical virtue, this argument would be unlikely to be compelling. I contend that unless one is an instrumentalist with respect to these models, there are theoretical reasons to prefer the HF model which trump the pragmatic virtues of the BF model.

4. Measurement Invariance and Unidimensionality

With respect to measurement invariance, Chen *et al*. argued that the BF model is more suitable to investigate measurement invariance with respect to, say, group (e.g., sex; we provide the

definition of measurement invariance below). Specifically, Chen *et al.* argued that the BF model enables the comparison of mean group differences in the general factor g^* and residual group factors ζ^*, whereas in the HF model measurement invariance is supposedly testable with respect to only the second order factor, *g*. However, there are two features of the BF model that make this claim problematic. First, indicators in the BF model are inherently multidimensional. Second, contrary to what Chen *et al*. argue, measurement invariance *can be* and often *is* tested with respect to group factors in the HF model (see, for example, Byrne & Stewart, 2006; Wicherts, *et al.*, 2004, 2005; van der Sluis, *et al*., 2007). The portion of the HF model that relates group factors to indicators specifies a unidimensional measurement model (a model which posits that differences in indicators is the result of only one common cause), and thus can be tested for measurement invariance.

Consider a measurement model in which indicators **y** are causally influenced by a latent variable η. This could be any psychometric measurement model (including various IRT models; Mellenbergh, 1994a), but I will focus on the linear factor model as a measurement model (Mellenbergh, 1994b; Ferrando, 2002), as depicted in Figure 4.4.

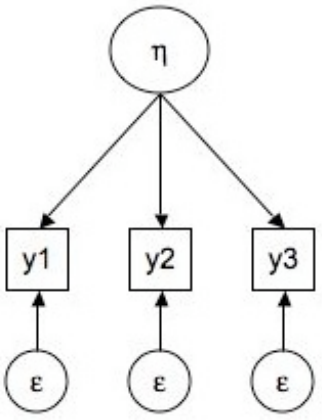

Figure 4.4: Psychometric measurement model. There are three indicators **y** and associated residual terms ε. η is the hypothesized measured ability (e.g., working memory).

In Figure 4.4 'η' refers to a cognitive ability, which I assume has a clear empirically or theoretically based definition, in the sense that the indicators were actually designed to measure, or established as measures of, this cognitive ability. Suppose that one, in accepting this measurement model, is a realist regarding η (Borsboom, Mellenbergh, & van Heerden, 2003). That is, η taken to be a mere statistical abstraction or linear combination of the indicators; rather, it as a hypothesized causally efficacious psychological (latent) trait that explains and causes variability in the indicator scores. Additionally, suppose one adopts the following psychometric criteria, which are independent of the issue of realism, but certainly desirable from a practical (statistical testing) point of view: the unidimensionality of the indicators, local independence, and measurement invariance with respect to the exogenous variable *V*. *V* may be continuous or discrete (Mellenbergh, 1989). For instance in the discrete case, *V* may refer to ethnic group, or sex. Measurement invariance implies that for all measurement outcomes Y, $f[Y|\eta^{o}]=f[Y|\eta^{o},V^{o}]$, where f is a link function, and the symbol '°' indicates a given fixed value of η and *V*. For example, consider Figure 4.5:

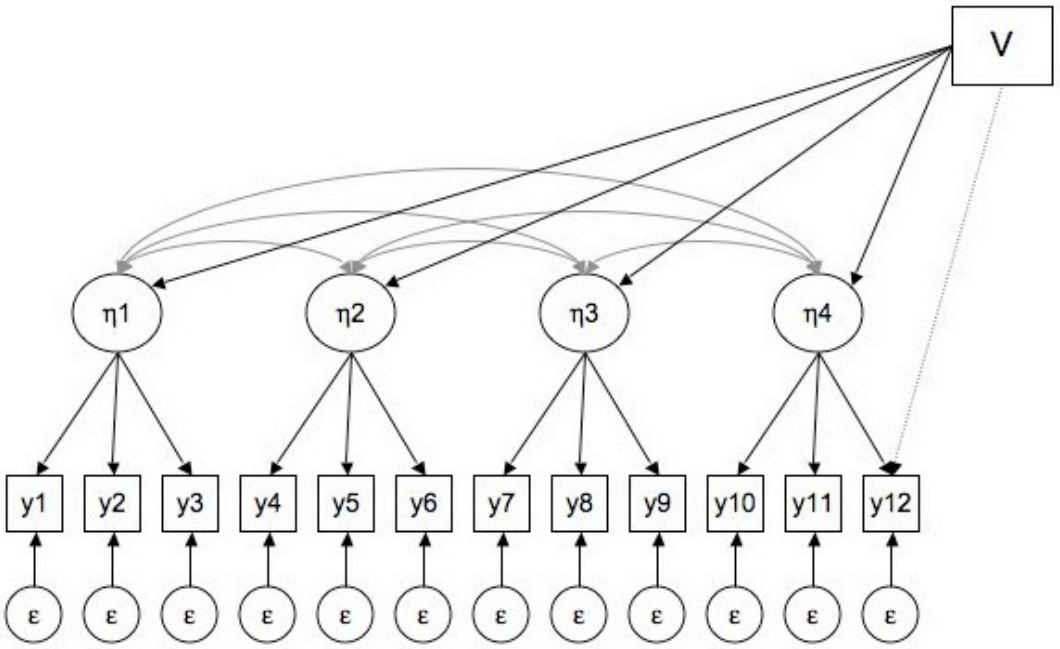

Figure 4.5: Measurement invariance and item bias. V is an exogenous variable that may be either continuous or discrete. The dotted line from V to y12 indicates that y12 is biased with respect to V. y1-y11 are not biased with respect to V since the relationship between V and $\mathbf{y}1-\mathbf{y}11$ runs via the latent variables $\boldsymbol{\eta}1-\boldsymbol{\eta}4$.

In Figure 4.5, the indicator $\mathbf{y}12$ is not measurement invariant with respect to *V*, i.e., the indicator, which might be a particular item or subtest, is biased with respect to *V*. If we attend specifically to the measurement model for $\boldsymbol{\eta}4$ the may be a bit more obvious. Measurement invariance with respect to *V* is testable and, in fact, it is a primary *desideratum* of item response models. See Meredith (1993) for a discussion of measurement invariance in the linear factor model.

I will now consider the BF and HF models in the light of realism with respect to the measurement model. First, with respect to unidimensionality, note that generally the BF model precludes the unidimensionality of indicators. To see this, attend to just that part of the measurement model that relates some group factor as well as g^* to their indicators in Figure 4.3. Interpreted as a measurement model, the BF model implies that test scores are indicators of both g^* and the residuals, $\boldsymbol{\zeta}^*$. Thus, under the assumption that the BF model is true, it is structurally impossible that one should succeed in constructing a unidimensional test for g^* or the group

common factor (except for the special case where the residual group factor has zero variance). This is a remarkable and counterintuitive feature of the BF model, which does not sit well with the psychometric notion of tests designed to measure a well-defined cognitive ability, and that test constructors may actually *succeed* in doing so.

A direct consequence of the two-dimensionality of the bifactor model is that, by necessity, there are no measurement invariant tests of general intelligence (only) or any other of the cognitive abilities in isolation. Naturally, measurement invariance *may* hold for the two-dimensional latent space defined by g^* and the specific factor jointly with respect to an external grouping variable V. This will be the case whenever the probability of a measurement outcome Y, given g^* and ζ^* is identical to the probability of Y given g^*,ζ^*, and V, i.e., $f[Y|g^*,\zeta^*]= f[Y|g^*,\zeta^*,V]$. However, the fact that test scores are the result of the influence of *both* g^* and the relevant group factor, entail that $f[Y|g^*]\neq f[Y|g^*,\zeta^*]$ and $f[Y|\zeta^*]\neq f[Y|\zeta^*,g^*]$, that is, measurement invariance with respect to g^* must be violated if one takes the test specific factor to be the group variable, and measurement invariance with respect to the test specific factor must be violated if one takes g^* to be the group variable The point here is not that forcing g^* or ζ^* into the role of grouping variable is sensible or not; the point is that, strictly taken, the BF model structurally precludes measurement invariance with respect to either factor in isolation. This is a structural feature of the model which holds independently of the battery of tests or items.

There are further consequences. To the extent that g^* is correlated with group membership (which is to say, to the extent that g^* differs over groups V=v), there can be no measurement invariant measures of any ζ^* (i.e., putative cognitive ability) in isolation. For instance, even if one assumes that the same model is true in both groups, i.e., that $f[Y|g^*,\zeta^*]=f[Y|g^*,\zeta^*,V]$, under the assumption that V and g^* are correlated we get $f[Y|\zeta^*]\neq f[Y|\zeta^*,V]$. Loosely speaking, V transmits effects of g^* to the observed score Y. Analogously, to the extent that ζ^* is correlated with group membership, there can be no

measurement invariant measures of g^* in isolation. That is, $f[Y|g^*] \neq f[Y|g^*, V]$. These relationships are diagrammatically represented in Figure 4.6 and Figure 4.7. For example, suppose Y is scoring in the top 25% in the upcoming bowling tournament, g^* is bowling ability, and V is membership in the Professional Bowlers Association (PBA). Since ζ^* must be uncorrelated with g^*, let us suppose that ζ^* is some factor that affects scores, but is (somehow) uncorrelated with bowling ability; perhaps ζ^* is how recently the player's lane has been conditioned at the time of the tournament. The choice of ζ^* may seem strange, but it is not clear what would be a suitable candidate for ζ^*; the situation is no different in the case of intelligence. It is plausible to suppose that g^* and V are correlated since one requirement for joining the PBA is that one have a bowling average of 200 or better for the most recent league season provided that there were at least 36 games in that season. Hence, the probability that one will attain a certain ranking in the upcoming bowling tournament given ζ^* is not identical to the probability that one will get that same ranking, given that one is a member of the PBA.

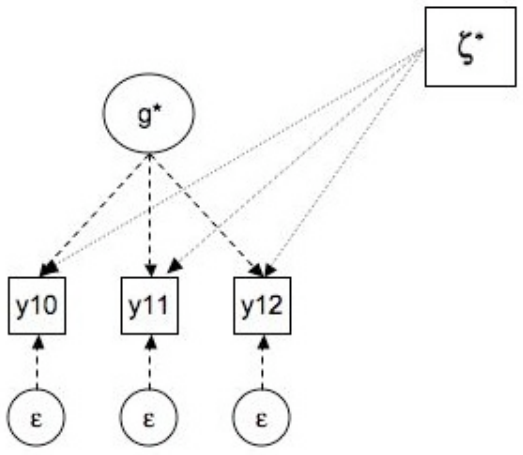

Figure 4.6: Violation of measurement invariance of **y**10–**y**12 with respect to the residual group factor ζ^*. That is, the indicators are biased with respect to ζ^*. Dashed lines indicate that the relationship need not be linear. The dotted lines from ζ^* (which is the violator) to the indicators indicate bias with respect to ζ^*.

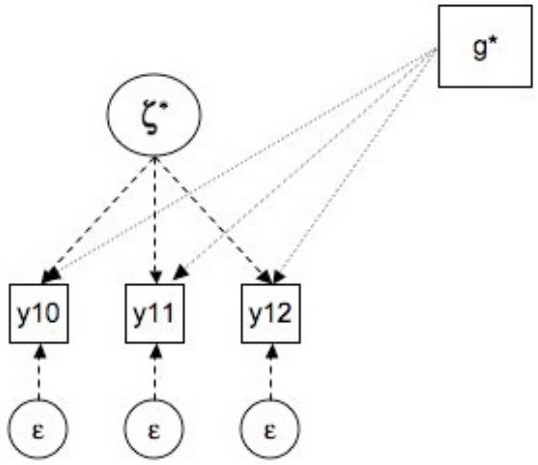

Figure 4.7: Violation of measurement invariance of **y**10–**y**12 with respect to g^*. That is the indicators are biased with respect to g^*. Dashed lines indicate that the relationship need not be linear. As before, the dotted lines from g^* to the indicators indicate bias with respect to g^*.

These observations are trivial psychometric implications of multidimensionality and should not be considered inherently problematic; that is, they should be accepted by any theoretician with the conviction that a set of test scores is inherently multidimensional, as the BF model assumes to be the case for intelligence test scores. The point is basically this: due to the multidimensional structure of the BF model, the factor g^* does not screen off effects of ζ^* and, as a result, it does not screen off any effects on the test scores that are conveyed via ζ^*; the converse also holds because ζ^* does not screen off effects that run via g^*.

Though these consequences may not be inherently objectionable, what does seem to be objectionable is that these implications are logically necessary consequences of the BF model; they follow from structural features of the model that are stipulated *a priori* and without substantive justification. As a consequence, if one believes the BF model to be true, and holds that the general factor of intelligence is g^*, then one is required, on pain of contradiction, to also believe that there cannot, in principle, be unidimensional measures of general intelligence. Therefore, one can set up a wide range of variables with respect to which g^* cannot be measured

in a measurement invariant way, some of which were spelled out above. However, measurement invariance and specific dimensionality are, arguably, *contingent* features of test scores; they are psychometric properties indiscernible *a priori*, which is why methodologists have developed testing procedures to find out whether they hold. Under the BF model, all such tests are rendered otiose merely in virtue of the model's formal structure. More importantly, there appears to be no substantive psychological *reason* why measurement invariant or unidimensional measures of general intelligence could not exist, and as far as we can tell there are several tests for specific abilities measured in standard IQ-test batteries that *do* appear to be unidimensional (e.g., inductive reasoning; de Koning, *et al*., 2003). A theoretician who holds that the BF model gives the proper statistical translation of the theory of general intelligence either has to challenge such empirical findings, or has to show that they are in some way artifacts.

5. Theoretical Considerations Pertaining to General and (Residual) Group Factors

In addition to Gignac and Chen *et al*.'s pragmatic and psychometric modeling arguments, conceptual reasons for preferring the BF model have been advanced as well. These considerations concern the theoretical status of the general common factor in the model. I will consider two kinds of theoretical considerations that are relevant to the interpretation of general and residual group factors. First I address whether g^* satisfies Jensen's theoretical constraints on general intelligence; that is, does g^* satisfy the constraints imposed by Jensen's theory. I argue that it does not and that g^* is not an adequate representation of general intelligence on this theory. The question then arises, given the statistical characteristics of g^*, what constraints does *it* pose on a theory of general intelligence? I argue that these constraints are unlikely to be met by a substantive psychological theory of intelligence. I then turn to a more general discussion of the identity of the factors on the BF and HF models and give further arguments for the claim that the factors in the BF model are not identical to those in the HF model.

6. Theoretical Constraints on the Interpretation of g^*

Carroll (1997) argued that in addition to satisfying statistical *desiderata* (such as goodness of fit), a model of general intelligence should also comport with substantive psychological theory. It is one thing to identify a general common factor, but it is another thing to demonstrate that factor (putatively) refers to a cognitive ability. Carroll writes (*ibid.,* p. 144),

> In any case, it would desirable to show also that general factor so identified constitutes a true ability, independent of lower-order factors, rather than being merely a measure of associations among those lower-order factors that might be due, for example, to the effects of common learnings.

This passage suggests a preference for the BF model if by "independent" Carroll means statistically independent. Carroll notes that IRT may be one possible approach to the identification of the general factor as a true ability but that within the context of the HF model this approach is problematic since

> …when a factor embraces tests that have multiple sources of variance, as would often happen, particularly if one is trying to demonstrate that a second-order factor, or worse still, a third-order factor, is a "true ability" independent of lower-order factors or variables…[,] items selected from measures of a higher order factor are likely to entail confounding with variance of lower-order variables, (1997, p. 145).

Note that in the context of the BF model these worries are supposedly ameliorated by the fact that the general factor g^* is specified as orthogonal to (i.e., uncorrelated with) the other residual group factors, ζ^*. Variability in indicators due to variability in the general factor ought not to be obscured by variability in indicators due to other residual first order common factors (though in a previous section we have argued that in general the BF model does not allow for this). Finding a general factor with these properties is an essential step toward a conception of a general factor as a well-defined, true (cognitive) ability, but it is only part of the story. Carroll admitted in 1997, and it remains true at the time this is being written, that "[w]hat has not been adequately

demonstrated and proven at this time is that *g* is a 'true ability' independent of more specific cognitive abilities defined by various types of psychological tests and observations," (*ibid.*, p. 151). I will argue that there are considerable, though not necessarily insurmountable, conceptual obstacles to a conception of g^* *qua* general intelligence within the BF model, Carroll's worries about IRT notwithstanding. It is worth noting that it is part of the essential nature of general intelligence, according to Jensen, that it does *not* satisfy the second *desideratum* in the latter quotation from Carroll, *viz.*, *g* is not characterized (or defined) in terms of specific tests, their contents, or, in the case of the HF model, first order abilities.

I have argued that *g*, as specified in the HF model, and g^*, as specified in the BF model, cannot refer to the same attribute unless a specific set of constraints holds (i.e., $\mathbf{\Lambda}_1 = \mathbf{\Lambda\Gamma}$), and that the models have very different psychometric implications attached to them. Therefore we cannot take *g* and g^* to be theoretically exchangeable representations of general intelligence (i.e., the construct hypothesized in substantive theories of intelligence). The question then becomes whether *g* or g^* is the proper representation of general intelligence as the construct has been described in the intelligence literature.

To answer this question we need a set of properties that general intelligence is supposed to have according to proponents of the relevant theory, so that we can match this list against the statistical and psychometric properties of the HF and BF models as discussed in the previous sections. Naturally, it is conceivable that every theorist has a slightly different conceptualization of what general intelligence is (Neisser *et al*., 1996); hence the results of the analysis may vary across such conceptualizations. In the present analysis, I focus on Jensen's account of general intelligence. There are two reasons for this: first, Jensen is particularly clear on what he takes general intelligence to be, and, second, he is one of the most important proponents of the theory.

Jensen's conceptualization of general intelligence is best conveyed in his own words (all quotes are taken from Jensen (1998):

> "Unlike group factors, *g* cannot be described in terms of the superficial characteristics of information content of the tests on which it is loaded" (p. 91).
>
> "The fact that a certain class of tests measures g more efficiently than other tests does not qualify the characteristics of the former tests to be considered the "essence" or "defining characteristics" of *g*. …The salient characteristics of the most highly g-loaded tests are not essential or definitional, but are empirical phenomena in need of theoretical explanation in their own right" (p. 92).
>
> "It is important to understand that *g* is *not* a mental or cognitive process or one of the operating principles of the mind, such as perception, learning, or memory. Every kind of cognitive performance depends up on the operation of some integrated set of processes in the brain. …But g is not these operations themselves" (pp. 94-95).
>
> "At the level of causality, *g* is perhaps best regarded as a source of variance in performance associated with individual differences in the speed or efficiency of the neural processes that affect the kind of behavior called mental abilities" (p. 74).
>
> "… *g* is not really an ability at all, but a property of the mind" (p. 92).

So, according to Jensen, the general intelligence factor cannot be described in terms of the features of a test's content, it is not a feature of the brain's circuitry (e.g., it is not a cognitive module; see Borsboom & Dolan, 2006), nor is it a cognitive process. Its nature is not to be discovered by examining the features of highly *g*-loaded tests, and the general intelligence factor is best regarded as a source of variance in performance associated with quantitative aspects of the neural processes that underlie cognitive activities; however, this factor is not itself a cognitive activity. Nevertheless, *g* is a psychological property, and that it is detectable is evidence that it is a causally efficacious one since epiphenomenal properties cannot effect observable differences at all.

It thus appears to be essential to Jensen's conception of *g* that this construct denotes some general source of individual differences, rooted in our different physical constitutions, that simultaneously influences individual differences in all the abilities that are measured by intelligence subtests. That is, the influence of *g* on test scores is *essentially indirect*. This feature

is clearly coded in the formal structure of the HF model; it is clearly violated in the formal structure of the BF model. Perhaps this is why Jensen claims that "[a]mong the various methods of factor analysis that do not mathematically preclude the appearance of *g* when it is actually latent in the correlation matrix, a [higher order] model is generally the most satisfactory, both theoretically and statistically" (*ibid*., p. 73).

Additionally, the two-dimensionality of the indicators in the BF model presents problems for the interpretation of g^* as a statistical representation of general intelligence. Notice that in the BF model, g^* is accorded the same status as the group factors, albeit more general, with respect to the indicators. So if one is a realist, g^*purportedly refers to some general cognitive ability. But recall the quotations from Jensen above: Jensen's *g* is *not* a cognitive ability, process, module, etc. Therefore, the bifactorists' g^* cannot be Jensen's (i.e., the higher order factorists') general intelligence factor even if g^* is regarded as a real or "true" ability. In other words, treating the general factor as designating a cognitive ability is inconsistent with Jensen's conception of the general intelligence factor. Hence, the bifactorists, insofar as they are realists, seem to be in need of a theory of intelligence.[45] This is especially clear in Carroll's case, for he wants to identify the general factor with a "true ability." It is incumbent upon the bifactorist to argue that g^* has the same psychological significance as the group factors; one may not simply presume that it does.

[45] Advocating realism with respect to theoretical entities *sans* theoretical commitments is a position known as "entity realism" in the philosophy of science and is formulated in Ian Hacking's (1983) *Representing and Intervening* (*cf.* Cartwright, 1983); however, this position has come under considerable scrutiny and it is not clear that it is a viable position for the bifactorist. For a discussion of the shortcomings of entity realism in the context of psychological measurement, see Trout (1998). For a more general criticism of entity realism's failure to make sense of scientific practice, see Leplin (1997, pp. 170–171 fn. 37; 2005). For criticisms that entity realism is ultimately uninformative because it prohibits ascribing causal properties to theoretical entities (which is precisely what realists about cognitive abilities do), see Psillos (1999, pp. 256–258). For this latter reason and because Hacking's realism is grounded in experimental intervention and exploiting the causal powers of theoretical entities to investigate more theoretical parts of nature, it is not straightforwardly applicable to the present concerns.

7. The Identity of Factors Across Models

Furthermore, contrary to what Chen *et al.* say, general intelligence does *not* "correspond to the general factor in the bifactor model," (Chen *et al*., 2006, p. 195) if the general factor in the HF model is understood in Jensen's sense. This is because the first order factors, of which g^* is one, *are* understood in terms of content of those test that load on them. This renders the interpretation of the residual group factor problematic, and likewise for g^*. If, however, the bifactorist insists, *modulo* any other concerns, that $g^*=g$, he is committed to a very strong position. To assert that $g^*=g$ is tantamount to asserting that HF=BF; the variables are identical only when the models are statistically equivalent. Now, this is somewhat ironic because any credence granted to the BF model will *ipso facto* be granted to the HF model. More importantly, asserting that the two models are identical commits one to the claim that the conditions under which the nested model is identical to the higher order model obtain, and this claim requires independent corroboration (not given by the BF model's apologists cited herein).

Nevertheless, we are not completely in the dark with respect to g^*. Some of its essential characteristics can be read off the model, though anything approaching a fleshed out theoretical account of g^* and its characteristics awaits the articulation of a concomitant theory. Nevertheless, spelling out the minimal constraints imposed on g^* by the BF model is sufficient to show that $g\neq g^*$, and, consequently, that the group factors in the HF model are theoretically distinct from the residual group factors in the BF model. To see why this is so suppose that g^* maps onto some property of the mind, P, with respect to which individuals differ. For example, if P is neural density, then differences in position on g^* will correspond to differences in neural density. For the purpose of this argument, the identity of P is inconsequential so long as there are quantitative individual differences in P-exemplification that correspond to position on g^*. Further suppose that the residual group factors are interpreted in the BF model as the group factors are in the HF model, i.e., as domain specific, psychometrically well-defined cognitive abilities. Gignac seems

to go along with such a supposition. For instance, Gignac (2005a, p. 324) interprets the residual group factors in the BF model as "verbal comprehension," "perceptual organization," and "freedom from distractability." Group factors in the HF model (*ibid*., p. 323) receive the same interpretation. Similarly, in his analysis of the Multidimensional Aptitude Battery, Gignac interprets the residual group factors in the BF model as "Verbal Intelligence," and "Performance Intelligence," while group factors in the HF model receive the same interpretation (Gignac 2006, p. 140). Chen *et al.*, too, in their analysis of models of quality of life also presumes that residual group factors in the BF model and group factors are co-referential (Chen, 2006, pp. 191, 193). Imagine we have a two populations of examinees, one representing a high position on g^* and the other representing a comparatively low position on g^*. Thus, each group is homogeneous with respect to g^* and, hence, P. We should expect to find no significant difference in terms of the residual group factors between this group and the population from which it is drawn, which is to say that the BF model predicts that, for example level of Verbal Intelligence, is indifferent with respect to g^*. Consequently, P is causally isolated from *any* of the residual group factors. This follows from the fact that all factors are orthogonal in the BF model, and that while correlation may not be sufficient for causation, it is certainly necessary. It seems quite unlikely that there could be such a property P underlying g^*; it is a property that is related to all tests of cognitive ability, but which is causally isolated from any particular cognitive ability. The statistical characteristics of g^* make it sound more like systematic error than a property of the mind. Thus, the bifactorist has three alternatives: give up interpreting residual group factors as group factors in the HF model are interpreted, give up interpreting g^* as a property of the mind, or give up both interpreting residual group factors as group factors in the HF model are interpreted, and interpreting g^* as a property of the mind. Any one of these options is inconsistent both with the idea that the factors in the BF and HF model are essentially the same and with Jensen's theory.

These interpretational problems do not arise in the case of the IRT model. To the extent that the model exhibits local independence, the measures are also unidimensional. Thus, η is well

defined (e.g., Working Memory). Once the empirical work as been done and the model has been fitted, there is no *a priori* obstacle to interpreting η. This does not mean, necessarily, that we are any closer to an understanding of the psychological significance or meaning of the general intelligence factor, but the HF model with *g* is no worse off than the BF model with g^* in that regard. Where the HF model excels beyond the BF model is in its explanatory resources vis-à-vis the group factors. Put simply, the HF model leaves us with the mystery of how to interpret *g*, but the BF model leaves us with the mystery of how to interpret g^* and an additional mystery, *viz.*, the interpretation of the residual group factors, for they cannot represent well-defined cognitive abilities. Viewed in this light, I do not agree with Gignac (2005, p. 312) that the BF model specifies the same "essential nature of factors," (Gignac, 2005, p. 321). Thus, awaiting its confederate theoretical apparatus, it seems that the most appropriate philosophical attitude regarding the BF model and its posits is instrumentalism. The model excels as a tool for describing the data, though it admits of no readily available theoretical interpretation. The concern is that waiting for the BF model's associated psychological theory may be, in at least one sense, like waiting for Godot.

Thus far I have pointed out theoretical dissimilarities between that *g* and g^*; they are essentially different with respect to their relational properties vis-à-vis the indicators and the (residual) group factors. I will now turn to *g* and g^* and argue that because the former is a second order latent variable and the latter is a first order latent variable, the two also differ with respect to their epistemic properties, despite any superficial, shared characteristics. For the purposes of this discussion, 'epistemic properties' will refer to qualities of *g* and g^* relevant to how we come to know that they exist as well as what evidence bears on such considerations. The discussion will be general, restricted to the status of *g* and g^* as inferred entities. Not all latent variables are on a par epistemically.

Consider g^* first. g^* is postulated to capture the variance of the indicators. It is an inferred entity that is but once removed from the observable indicators. If the factor model is interpreted as a measurement model, then *g* is a common cause of the variability in indicators (e.g., items scores or subtest scores). g^* is hypothesized to have direct effects on the indicators, ignoring mediating effects not included in the model. We confirm that g^* bears the relevant relationship to the indicators by fitting the appropriate measurement model to intelligence data (momentarily leaving aside worries about multidimensionality). The existence of g^* is, in principle, confirmable in the same way as group factors in the HF model.

g however, does not share these epistemic properties. First it is twice removed from experience. Its existence is inferred not from statistical properties of observed measures; rather, the existence of *g* is inferred on the basis of statistical features of posits that are, themselves, inferred, *viz*. group factors. Thus in the HF model, we can be no more confident in the assertion that *g* exists than we are that the group factors exist. In this sense *g* is epistemically parasitic upon the group factors; its evidential basis comes via the group factors. Conceiving of the HF model in terms of a measurement model, the well-defined cognitive abilities are the common causes of individual differences in the indicators, and *g* is postulated as the common cause of those first order common causes. That is, *g*'s place in the etiology of item scores is temporally prior to that of the group factors, thereby placing it at a greater inferential distance from the indicators than the group factors. In the BF model, however, the epistemic (inferential) gap between indicators and g^* is no different than the gap between indicators and residual group factors; they are both purported to be measured abilities once removed from experience. Furthermore, the epistemic status of g^* is independent of the epistemic status of the residual group factors. Thus it is evident that *g* and g^* do not share their epistemic properties, even under the assumption proportionality constraint obtains. Note that the differential epistemic status accorded to *g* and g^* does not entail that the two are distinct entities. It may, nevertheless, turn out that *g* and g^* pick out the same features of cognition despite differing with respect to their epistemic properties. This is to say that

empirical investigation may reveal that the two are but nominally distinct (though I have given reasons to think this would be extremely unlikely).

There is a further way in which questions of realism bear on the interpretation of a model, particularly in the context of how the factors in the HF model are interpreted. On one interpretation of the HF model, which may be aptly called a "conservative realist," conceiving of *g* as an ability is suspect. On the other interpretation, which may be aptly called a "liberal realist," conceiving of *g* as an ability is permissible and, indeed, in line with how *g* is considered by some (e.g., Carroll, 1997). Thus, for the conservative realist the first order factors are subject to measurement, but the higher order factor is not. Therefore, evidence for the existence of a higher order factor cannot come from fitting a measurement model alone. Rather, the relevant evidence is found in the explanatory merits of the structural model that explains the structure of correlations among the first order factors. For the liberal realist, the structural part of the model (i.e., the relation between the higher order factor and the first order factors) is also considered to specify a measurement relation, be it an indirect one. Evidence for the *g* factor can come from fitting a measurement model alone, because in this interpretation the relation between the higher and lower order factors is part of the measurement model. We will address these interpretations in turn.

The concern of the conservative realist regarding higher order factor models is this. To interpret *g* as an ability is to make an unwarranted inferential leap from the HF model to cognitive systems. The inference is dubious, because it conflates the part of the HF model that is a measurement model, i.e., the part that relates group factors to indicators, with the part of the model that is a structural equation model, i.e., the part that relates *g* to the group factors (*cf.* van der Sluis, *et al*., 2006). Conservative realism makes agnosticism the default position regarding higher order factors. There are advantages to the conservative realist position. Attending to this distinction between a measurement model and structural equation model leaves one with considerable interpretational room regarding *g*. Of course, it is consistent with the HF model that

g is a "higher order" cognitive ability (though it is inconsistent with the Jensenist line), but, according to the conservative realist, this is not a fact that one should read off of the model, for it ignores an important methodological distinction that may block interpreting *g* as an ability. Such an interpretation of *g* would command considerable theoretical work since the relationship between the indicators and *g* is *not* one of measurement, according to the conservative realist. It is an especially attractive feature of this distinction that it is consistent with, if not recommendatory of, theoretical economy with respect to *g*. While the model may commit us to the existence of cognitive abilities, such as WM, etc., it does not commit us to the existence of a general cognitive ability, for if we take the model seriously, measurement occurs only at the level of the measurement model. According to conservative realism, it makes no sense to speak of "measuring *g*."

Conservative realism, a metaphysically conservative position with respect to *g*, has further interesting consequences. For example, it suggests and rationalizes a methodological stance when a group factor such as WM is identified with *g* (Kyllonen, 2002; Kyllonen & Christal, 1990; Colom *et al*., 2004; Gustaffson). Such a situation suggests two alternatives: 1. interpret the correlation to be evidence that one has "found" *g*, i.e., that the essence of *g* has been captured, or 2. revert to the oblique first factor model depicted in Figure 4.1, without treating *g* as an independent ability or existing over and above the group factors. The latter alternative is one that conservative realism recommends.

In such a situation, taking the metaphysically conservative position requires that we not multiply entities beyond necessity and, therefore, we should revert to the oblique first order factor model *à la* Figure 4.1. This is the natural choice if one regards the *g* factor as a mere mathematical transformation or if one toes Jensen's line in saying that *g* is not a measured ability. The HF model does not force *g* into the role of a psychometric ability as it does with the first order factors; furthermore, this is consistent with taking a realist stance with respect to the measurement model. If one takes it to be a salient point of designation that in the HF model *g* is

part of a structural equation model (i.e., not a measurement model), then the fact that a correlation of 1 between *g* and, say, Working Memory (WM) obtains seem to indicate that the two are only nominally distinct, i.e., that there is but one entity by two names, thereby obviating the necessity of positing two entities. That is, if we can do all the explanatory work without postulating *g*, something whose nature is unclear and which still stand in a mysterious relation to the indicators that do not measure WM, then methodological prudence would seem to dictate that we resist any inclination to compound our ontological commitments. Note that this situation also points out the metaphysical excesses of the bifactor model, which by necessity posits both a general factor and group factors. In sum, if a realist about latent variables draws a principled distinction between latent variables in a measurement model and latent variables in the structural equation model, then there is room to restrict one's ontological commitments to first order factors. The bifactorist, however, cannot exploit this distinction in his model; if he is a realist, he is a realist about g^* and the other first order factors. But as suggested earlier, the realist position is not open to the bifactorist since his model is purely statistical, i.e., it suffers from a dearth of theoretical resources on which to draw for the interpretation of its theoretical posits.

The liberal realist, on the other hand, does not believe that the distinction lying at the core of the conservative realist position does anything but point out parts of the model that are merely nominally distinct. That is, the liberal realist treats both the top half of the model and the bottom half of the model symmetrically; both are measurement models. The HF model represents a causal chain running from indicators to *g*. Though perhaps not generally, liberal realism accepts the transitivity of causation, for if an indicator measures a cognitive ability, and the cognitive abilities are measures of general ability, then the indicators, too, measure general ability, though indirectly. One advantage of this position over the conservative realist position is that it does not invoke, and therefore need to defend, the asymmetry with respect to the interpretation of latent variables at different levels of the model. But on the other side of the coin, liberal realism is more metaphysically ostentatious and is, thus, more epistemologically risky. It is also, arguably,

inconsistent with Jensen's conception of the *g* factor. This is not itself an objection, but recall that it was shortcoming of the BF model that it did not have an associated theory of intelligence to make sense of the posited abilities, unlike Jensen's *g*. The liberal realist is in a similar situation. Not only must he come up with a concomitant theory of intelligence, but he must also answer to Jensen's claims about the fundamental nature of *g* as well as to those who have shown that the data can be explained without positing a general factor (*cf.* van der Maas *et al.*, 2006). Toward addressing the former concern, the liberal realist could attempt to construct a battery of tests that load perfectly on first order factors. If successful, he will have rendered each first order latent variable a manifest indicator of *g*, in effect, and empirically corroborated the claim that measures of first order factors are also measures of the second order *g*. However, this does little in the way of addressing the latter concern.

Where the two positions may agree, however, is in their response to finding a first order factor that correlates perfectly with *g*. The considerations for theoretical economy advocated by the conservative realist may also be invoked by the liberal realist, for presumably it is because *g* is said to do explanatory work in the HF model that it is posited at all; however, if *g* can be exorcized from the model without sacrificing explanatory power, then, presumably, the liberal realist would be able to abandon *g* as a theoretical posit. It should be noted, however, that if theoretical economy were foremost among the virtues that one requires of his models, then it is not clear that there would be much room for *g* at all, especially if the HF model and the oblique first order factor model fit equally well, as Gignac (2005, p. 325) reports.

8. Conclusion

The BF model provides a better fit to intelligence data and the quality of life data considered by Chen *et al.*. Additionally, the bifactor model is alleged to have other desirable psychometric properties that purportedly make it a preferable alternative to the HF model. Carroll suggests that the BF model facilitates conceiving of 'g^*' as denoting a "true ability." That is, the interpretation

of g^* is more straightforward in the context of the BF model than is the interpretation of *g* (which he and others conflate with g^*) in the HF model. However, as Carroll notes, a concomitant psychological theory of general intelligence has yet to be articulated. I argued that in terms of measurement, the HF model has certain desirable properties not exemplified by the BF model. Assuming a realism regarding the measurement model for η (i.e., that first order factors purport to refer to real cognitive abilities), I argued that the HF model is no less attractive than the BF model. In fact, it is unclear that realism sits well with the BF model at all; though from an instrumentalist perspective, the BF model may be preferable in virtue of its better fit to data. However, this gain in empirical adequacy comes at the cost of explanatory power.

The HF model does not suffer the BF model's interpretational problems. First the HF model retains the measurement model for η and the associated interpretational and modeling advantages. For example, HF does not preclude unidimensionality of a set of indicators. Also, it is possible to test for measurement invariance of particular η-scores. Second, the HF model is consistent with a conception of *g* as a quantitative dimension underlying all cognitive abilities. Third, the HF model does not force *g* into the role of a psychometric latent variable, i.e., a measured cognitive ability.

According differential theoretical status to *g* and first order factors is not arbitrary and is well motivated given if *g* is not a *measured* cognitive ability (e.g.,Working Memory). This distinction between measured cognitive abilities and inferred entities such as *g* is particularly advantageous since *g*, at least as conceived by Jensen, is not *supposed* to be a cognitive ability, though Jensen does consider it to be a causally efficacious psychological property. To treat *g* as a cognitive ability, as Carroll does, would be to accord *g* the same status as η, and to interpret g^* in the same way as Jensen does *g*, as has been seen, invites a host of *prima facie* problems. That Carroll's interpretation of *g* is inconsistent with the Jensenist interpretation is not necessarily a problem provided a plausible alternative interpretation is offered. None is presently known, which

is not to downplay the instrumental virtues of the BF model, but the model does seem to be a step backward both in explaining intelligence data and in moving toward a theory of intelligence.

I have discussed the relative advantages and disadvantages of a bifactor vs. higher order factor model of intelligence, focusing primarily on psychometric considerations and considerations pertaining to the nature of *g* as a latent variable. Several important consequences follow from adopting the bifactor model in conjunction with realism regarding its theoretical posits that seem important to consider:

1. g^* is uncorrelated with other first order factors (typically taken to refer to cognitive abilities).
2. There can be no unidimensional measures of cognitive abilities or g^*.
3. There can be no measurement-invariant measures of g^* only or of any other cognitive ability in isolation.
4. To the extent that g^* differs over groups, there can be no measurement-invariant measures of any of the other cognitive abilities over these groups.
5. To the extent that cognitive abilities differ over groups, there can be no measurement-invariant measures of g^*.

I also noted that the BF model is not compatible with Jensen's theory of intelligence and conception of *g*. The BF model as of yet awaits a theory of intelligence that entails each of these statements. Its primary virtue is its relatively superior fit to data. For these reasons, the BF model is best regarded instrumentally, as a tool for describing or reducing data, i.e., not as representing individual differences in cognitive facility. This is to say that realism is not a tenable philosophical position for the bifactorist. However, this may change should an appropriate psychological theory come along. Finally, I suggested that there are two ways one can be a realist regarding the HF model, depending whether one interprets factors in the structural part of the

model analogously to how one interprets factors in the measurement model. The difference between these two ways rests largely on epistemic considerations.

CHAPTER FIVE

ON THE CAUSAL INTERPRETATION OF LATENT VARIABLES

As I sleep, some dream beguiles me, and suddenly I know I am dreaming. Then I think: This is a dream, a pure diversion of my will; and now that I have unlimited power, I am going to cause a tiger.

Jorge Luis Borges (1960) *Dreamtigers*

1. Introduction

Molenaar (1999; Molenaar, Huizenga, & Nesselroade (2003)), Jensen (1998), Deary (2002), and Borsboom (Borsboom *et al.,* 2003, 2005; Borsboom & Dolan, 2006), provide a technical reason for thinking that general intelligence cannot be construed as a causally efficacious attribute responsible for individual item responses. The reason, which will be explained in more detail below, is that the models of general intelligence are not *locally homogeneous*. Consequently, attributions of general intelligence to individuals are unintelligible.

I will argue for several claims:

1. Models of general intelligence are not necessarily locally homogeneous.

2. Psychological and psychometric practice reveals a presumption of local homogeneity.
3. If models of general intelligence were locally homogeneous, additional evidence would be required to establish that general intelligence is a within-subject attribute.
4. If models of general intelligence were not locally homogeneous (and they are not), additional evidence would be required to establish that general intelligence is a real between-subject property.
5. Conceptual considerations of the nature of general intelligence *prima facie* rule out the possibility of making sense of it as a within-subject attribute.

While others have pointed out the problems associated with assuming local homogeneity of latent variable models, little has been written on how those problems bear on general intelligence. Borsboom (2005) points out that general intelligence might be in trouble vis-à-vis considerations of local homogeneity. My goal here is to spell out the details of this concern. I then make the case that psychometricians who interpret the *g* factor causally do, if only tacitly, assume local homogeneity. I will discuss causal interpretations of *g* and psychometric abilities in general, and relate the discussion of local homogeneity to Borsboom's conception of validity in psychological testing.

2. The Local Homogeneity Assumption

Local homogeneity is a (contingent) relation between models of variability in a population (between-subject variability) and models of variability in elements of that population (within-subject variability). Within-subject models are typically constructed from *time-series* data. The relevance of time-series data in models of within-subject variability is that there must be repeated measures of the attribute being modeled in order to generate variability. Thus in time-series analyses, N=1 (subject drawn from a population **P)**, and the number of measurement occasions,

T, is large. Time-series data in the behavior sciences are typically used to generate dynamic factor models that track development. In a standard factor analysis of between-subject variability, N is usually large (and also drawn from **P**) and T is small. The factor structure of test score covariation for the individual over time are the same as the factor structure of the between-subject design only if the factorial structures are locally homogeneous. This follows from Ellis and van der Wollenberg's (1993) Theorem 3: Suppose *L* is a specified latent trait model. *L* holds with local homogeneity for **X** in **P** if and only if *L* holds for **X** in every positive subpopulation of **P**, where **X** is a set of observed score patterns. The case of N=1 is just a specific subpopulation of **P**. That is, "the homogeneity assumption implies that all subjects in a given population obey exactly the same factor model," (Molenaar *et al.*, 2003). The conditions required for local homogeneity are strong. The first condition is *stationarity*: each member of the ensemble must have stable statistical characteristics, such as a constant mean levels. This condition alone rules out locally homogeneous models of developmental processes since they have statistical characteristics that vary over time. The second condition is *homogeneity of the ensemble*. If the ensemble is homogeneous, the trajectories of each individual fall under the same dynamical laws. Local homogeneity, causally interpreted, is taken to imply that a given test measures the same attribute in every member of the population and every subpopulation. In other words, "[i]f one says that the test measures a certain trait for the subjects of **P**, then it should also measure that trait for, say, the women of **P**," (Ellis & van den Woollenberg, 1993). The causal interpretation, for reasons that I will discuss later, is problematic.

The unwarranted assumption of local homogeneity is not just some esoteric problem confined to psychometrics. *Ergotic theory*, a subdiscipline of statistical mechanics is concerned with the relationship between dynamical and thermodynamic systems, i.e., whether the methods of statistical mechanics can be used to study dynamical systems and the conditions required to

prove the identity of averages over time and averages of ensembles (Farquhar, 1964).[46] In statistical mechanics, the *ergodic hypothesis* is the claim that "in its motion through phase space, the point representing [a] system spends in each region a fraction of time proportional to the volume of the region," (Ruelle, 1991, p. 111). A proof of the ergodic hypothesis is still outstanding. The details of ergodic theory are not particularly important here, but note that at the general level, the kind of issue that concerned Gibbs and Boltzmann the same as that with which psychometricians are concerned: the relationship between time-series data (i.e., dynamical) and data describing ensembles, and how one goes between the two.

3. Models of General intelligence are not Necessarily Locally Homogeneous

Molenaar and others have conducted simulation studies aimed at showing that standard factor analyses of variation in populations are insensitive to within-subject heterogeneity. These simulation studies also show that local homogeneity is not built into latent trait models. I will describe two of these studies. The first is due to Molenaar (1997) and the second is due to Molenaar (1999). Following the description of the simulation studies, I will spell out the practical consequences of heterogeneity.

In the first simulation study there are N subjects, each of whose behavior is specified by a different factor structure (up to 4 factors). One subject may obey a 1-factor structure, another a 2-factor structure, and each subject is associated with different factor loadings and measurement-error variances. Thus with respect to within-subject variability, there is radical heterogeneity. The question, then, is whether there is a between-subject factor model that adequately describes the between-subject variability. If so, then local homogeneity is violated because not every member of the population could exemplify the between-subject model. Molenaar found that a 1-factor structure was sufficient to fit the between-subject variability. This is surprising because most

[46] It was Boltzmann who in 1871 introduced the term 'ergodic' to refer to systems that were isolated, classical, and whose phase trajectory passed through every point of the surface of constant energy in phase space that corresponds to the energy of a system (Farquhar, 1964, p. 75).

subjects' time-series data were (by construction) not fit by a 1-factor model and for those whose behavior was specified by a 1-factor model, the factor loadings and measurement-error variances of the between-subject analysis did not match those associated with the time-series data.

The second study does not differ importantly from the first simulation study. However, in the second study Molenaar also determines the factor scores for each subject on the basis of the between-subject model and correlated those scores with the factor scores derived from the time-series data. The correlations were low and in some cases negative.

As Molenaar, Huizenga, and Nesselroade (2003) point out, the consequences of these studies for psychology are substantial. Tests developed using between-subject data, while they may measure the source of *between-subject* variability adequately, may yet fail to measure the same thing in individuals within the population for which the test is intended.[47] For example, an intelligence or personality test developed using standard factor analysis may not measure the relevant construct at the level of individuals. In terms of the studies above, the satisfactory fit of a 1-factor model (for the between-subject data) suggests that the test is unidimensional, i. e., it measures one attribute only; however, for most of the subjects, a multidimensional model fit best (by construction). Moreover, that negative correlations arose between the results of the standard population-level analysis and time-series analyses seems to suggest that if the test were used for purposes of prediction, they would fail for most of the individuals in the population. These considerations confront a viable account of validity with an additional burden: an account of validity must be sensitive to the distinction between within-subject variability and between-subject variability lest we christen a test as valid on the basis of standard factor analytic studies when it measures something in individuals other than what it measures in populations. Now, it remains to be seen whether statistical models of intelligence (particularly those with a *g factor*) are locally homogeneous. However, since, in general, models may not be locally homogenous,

[47] Note here the assumption that if local homogeneity holds, the *sources* of between-subject variability and within-subject variability are the same.

models of intelligence may not be either. I will argue that models of general intelligence are not locally homogeneous and that the assumption of local homogeneity underpins the practice of intelligence testing.

Now it seems (to me at least) obvious that intelligence, mental ability, what have you, is usually taken to be an attribute of individuals. There is no shortage of competing theories of intelligence, but all mainstream theories (and even some of those outside the mainstream such Howard Gardner's theory of multiple intelligences (Gardner, 1983)) posit mental ability (or "intelligence") as a property of individuals. Also, we say things like "John did so well on the test because he's so intelligent" or "Look at how well little Jaime did on her math test; she's so intelligent." Of course, these folk psychological claims are typically completely divorced from substantive psychological theory, but nevertheless, they indicate a commitment to intelligence as some causally efficacious property of individuals. Moreover, these folk psychological claims are not that different from what one finds in a clinical report of one's performance on an IQ test. Therefore, I take it that intelligence is plausibly construed as psychological attribute within individuals. However, psychometric theories of mental ability are based on between-subject analyses of test performance. They have focused on (differences in) intelligence as a source of individual differences, i.e., differences in intelligence are posited to explain differential performance on tests of mental ability. The obvious and well-worn way to get to the individual from the population is *via* the assumption of local homogeneity, otherwise the tests may be measuring different traits in individuals than they are for the population. I will argue the *g* factor *cannot* be understood on the basis of between-subject data as denoting mental ability *qua* within-subject attribute.

Psychological practice seems to indicate that psychologists do assume local homogeneity, if only tacitly. The concept of intelligence on which the most popular intelligence tests are based has general intelligence as a central theoretical posit, and general intelligence has its provenance in standard factor analysis of population-level data, not time-series analyses of within-subject

variability. The commitments of psychometricians are difficult to discern. Famously, Spearman hypothesized that *g* was mental energy, a within-subject attribute. However, he also cautioned his readers that the *g* factor was only a statistical construct expressing between-subject variability. Jensen, too, does not seem consistent enough to attribute to him a commitment to local homogeneity. Consider the following quote from Jensen (1998, p. 95):

> It is important to understand that *g* is *not* a mental or cognitive process or one of the operating principles of the mind, such as perception, learning, or memory. Every kind of cognitive performance depends upon the operation of some integrated set of processes in the brain. These can be called cognitive processes, information processes, or neural processes. Presumably their operation involves many complex design features of the brain and its neural processes. But these features are not what *g* (or any other psychometric factor) is about. Rather, *g* only reflects some part of the *individual differences* in mental abilities…that undoubtedly depend on the operation of neural processes in the brain.

However in a series of interviews with Frank Miele (2002, pp.58-59) on the *g factor* and intelligence, Jensen refers to an individual's *g* as being causally relevant to determining that person future occupational success. Mike Anderson (1992, p. 2) indicates that he assumes local homogeneity when he writes that

> [s]ince differences in tests scores are the target of explanation, whether these represent differences between 2 adults or longitudinal changes within the same individual seems irrelevant. It is taken to be a parsimonious assumption that these differences in scores are to be explained with reference to the *same mechanism*. Thus, for example, higher synaptic efficiency makes on individual more intelligent than another, and increasing synaptic efficiency with age makes us more intelligent as we develop.[48]

Kanazawa (2004) also assumes local homogeneity when he hypothesizes that *g* is a species-typical information processing mechanism (*cf.* Borsboom & Dolan, 2006). But knowing who can be justifiably accused of assuming local homogeneity would seem to get us nowhere. The mere fact that the tests are taken to measure the same attribute in all subpopulations of the population for which the test is intended indicates a commitment to the assumption. Clearly in testing and

[48] Italics added.

marketing of tests, the assumption is made. Moreover, the construction of norm-referenced tests seems to presuppose local homogeneity, for it is assumed that the trait being measured in the standardization group (from which the norms are generated) is the same trait that the test measures in the population that the standardization group is intended to represent. The proof is in the practice if not in proclamation.

4. If Local Homogeneity Holds

Suppose that models of general intelligence are locally homogeneous. Thus we have the same factor structure representing the population at large as we do for every positive subpopulation, in particular every individual. The psychometrician is likely to interpret this feature of his models to mean that tests measure general intelligence in the population as well as in each individual; the source of between-subject differences is the same as that responsible for within-subject differences. This latter claim is extremely problematic, but I will bracket it momentarily. What I wish to point out is that local homogeneity is a contingent, structural relationship between statistical models. Any causal implications must be added, as etiology is not reported by the statistics; even when statistics are taken to indicate a nonspurious relationship, they still do not fix the direction of causation, for example. Local homogeneity may be *circumstantial* evidence that there is a common source of variance (i.e., attribute) shared by within-subject analyses and between-subject analyses of the variability in performance on some battery of tests, but it is not entailed by local homogeneity. That is, identifying a class of models as locally homogeneous does not absolve the psychologist or psychometrician from establishing 1.) that the test, in either research design, is measuring anything at all; the ability to extract a latent variable or fit a confirmatory model is not tantamount to discovering or confirming a real common cause; and 2.) that the relationship between the models is not otherwise spurious, e.g., showing that the latent variable in the between-subject design is the same or denotes the same attribute as the latent variable in the within-subject design. Thus, even if models of general intelligence *were*

(syntactically) locally homogeneous and we had good reason to believe that *g* was "real", that fact would not be sufficient to justify realism about general intelligence as a causal attribute at the level of individuals. We would still have to show that the sources of variability in each design are the same, their factor structures notwithstanding. Form does not fully specify content. Validity, which *is* a causal concept, is required for that. To reiterate: Molenaar's Theorem 3 proves a fact about the syntactic relationship between models; it is not equivalent to, nor does it imply, the claim that the attribute being measured in the population is identical to the attribute being measured in the individual. The latter claim is an unjustified empirical interpretation of a statistical property of latent variable models.

5. If Local Homogeneity Does Not Hold (and it doesn't)

Suppose that models of general intelligence are not locally homogeneous. Thus we have a factorial structure representing the population at large that differs from the factorial structure representing some subpopulation or other, including subpopulations of unit size. The situation is not that different from the situation where the assumption of local homogeneity holds. In particular, a failure of local homogeneity does not imply that the source of variability in the population is different from the source of variability in the relevant subpopulation. In some cases, such as in Molenaar's second simulation study, heterogeneity will indicate that the sources of variability are different, but there's no reason to think that heterogeneity alone implies this fact; the relationship between models and sources of variability is one of indication, not implication. It remains an open question whether the sources of variability in each model are identical, despite their resistance to be fit by the same model. It may vary depending on the theory of the trait being considered. A single source causal mechanism may show itself differently in a between-subject design than it does in the within-subject design.

Models of general intelligence are not locally homogeneous. The problem with general intelligence is that it is supposed to be a relatively stable attribute. More precisely, there is little

variation in scores across repeated measures for an individual. Typically, variation between testing occasions is attributed to measurement error, not variation in ability. Psychological theory and psychometric data tell us that mental ability is stable, but if it is, then there is no within-subject variability to model, i.e., no time-series analysis is available for the individual. With no variability, there is no latent variable to be extracted. At the population-level, however, we find that the *g* factor models are robust. As Jensen says in the quoted passage above, "*g* only reflects some part of the *individual differences* in mental abilities". Jensen (2002) makes a more careful statement relevant to the issue of local homogeneity in the context of intelligence research and psychometric models of individual differences:

> It is important to keep in mind the distinction between intelligence and *g*… . The psychology of intelligence could, at least in theory, be based on the study of one person, just as Ebbinghaus discovered some of the laws of learning and memory with N=1… . Intelligence is an open-ended category for all those mental processes we view as cognitive, such as stimulus apprehension, perception, attention, discrimination, generalization, learning and learning-set acquisition, short-term and long-term memory, inference, thinking, relational education, inductive and deductive reasoning, insight, problem solving, and language. The *g* factor is something else. *It could never have been discovered with N=1*, because it reflects individual differences in performance on tests or tasks that involve any one or more of the processes just referred to as intelligence (pp. 40-41).[49]

That is, *g* is a between-subject statistic, and what it purportedly denotes is a between-subject attribute that "explains" the positive manifold. The fact of heterogeneity does not imply that the between-subject source of variability is not also a source of variability within-subjects. Consider the attribute *height*. Height seems to be an attribute that explains both within-subject and between-subject variability on certain measures such as being able to ride a roller coaster, retrieving items from high shelves, weight (for adolescents), and shoe size. With general intelligence, however, all we have are between-subject models which tell us nothing about how the attribute functions in individuals, thus to make inferences about or on the basis an individual's

[49] Italics added.

"general intelligence" being a causal factor is, arguably, unwarranted. Individuals may have some attribute that we can identify as indicative of "intelligence", but the between-subject model does not tell us if it is the attribute purportedly indicated by the *g* factor, though those within-subject attributes may be related to general intelligence (but this relationship is not implied by the model).

The objection against assuming local homogeneity might be further elucidated by considering a couple of prosaic examples. 80% of the population of Tbilisi is Georgian, however it is not true of each resident of Tbilisi that he is 80% Georgian. Some residents of Tbilisi are Armenian while others are Russian. Even the Georgians are likely not to be 80% Georgian (by whatever metric). Here we have a lack of local homogeneity because the categorization scheme that fits the population does not fit the individual. The following example is even closer, structurally, to what is going on in the case of general intelligence where we have variability in the attribute only at the level of the ensemble, but none to be modeled in the individual. Heritability estimates purport to report the portion of phenotypic variance in a population that is attributable to genetic differences *in the population*. By adulthood, heritability estimates for height are greater than or equal to .9. This means that at least 90% of the total population variance in height is attributable to genetic differences between members of the population. However, it is a common misunderstanding of heritability estimates that they report some feature of individuals in the population, e.g., that for each (or any) adult in the population, his height is 90% genetic. Heritability estimates are undefined for the individual.

Thus far I have focused on heterogeneity of interindividual and intraindividual measurement models, but as suggested in the discussion of Theorem 3 above, heterogeneity can also obtain between models of populations and their subpopulations. Recall that if local homogeneity holds, then a measurement model that fits the ensemble should fit every positive subpopulation. Thus a test that measures general intelligence also measures intelligence in men.

Sometimes relationships between variables will remain measurement invariant across subpopulations, but this is not always the case. Simpson's Paradox makes this especially clear.

Simpson's Paradox's claim to fame (other than exonerating Berkeley University from charges of sex-discrimination (Bickel, *et al.*, 1975)) is the difficulty it poses for probabilistic theories of population-level causal claims. Instances of the paradox are characterized by statistical or probabilistic relationships between two variables that reverse or disappear in subpopulations. It is a paradox associated with reference class heterogeneity, which is precisely the problem the psychometrician encounters when a measurement model is not invariant across subpopulations, i.e., when they are nonergotic. For example, van der Sluis *et al.* (2006) analyzed sex differences in performance on the WISC-R in Dutch and Belgian samples. They found violations of measurement invariance with respect to sex in the Information, Arithmetic, and Coding subtests. Consequently group comparisons with respect to ability cannot be made legitimately on the basis of these subtests, for they measure different things in different groups. If the subpopulations were partitioned not only by sex, but some other (possibly spurious) variable, further heterogeneity may become evident. It is unlikely that once measurement invariance is found to be violated that one would be inclined to attempt to fit the model to the population as a whole since this property is typically interpreted as the presence of test (or item) bias with respect to some exogenous variable. I have focused on the relationship between individuals and the population since I am primarily interested in how *g* applies to the individual, if it applies at all; however, interesting problems come up when we consider group differences between subpopulations. I will not be addressing those problems here.

We have seen that models of general intelligence are not locally homogeneous. Interpreted causally, this means that the source of variability in test scores in a population are different from those operating within individuals who comprise the population. I have resisted the causal interpretation noting that when there is heterogeneity it serves as a reminder that additional evidence is needed to establish whether the same processes that explain differences between

members of the population also explain differences within the individual over time. Thus we cannot infer from the robustness of the *g* factor that it tells us anything about intelligence as a property of individuals. I have pointed out that the prospects are bleak for interpreting the *g* factor causally (within individuals). These interpretations may even be misconceived, for *g* just isn't the right kind of thing to fit that description. An account of intelligence and its causal role in individuals calls for an individual-specific statistic. Such a statistic would enable us to formulate coherent claims about the causal role of intelligence in individual behavior. The *g*-score is one such statistic. In what follows, I will consider one way of testing causal claims about individual intelligence and argue that there are both conceptual and practical reasons to think that the explanatory resources of the *g*-score are rather anemic. I will recommend that psychometricians curtail their epistemic aspirations, though my critique will stop short of eliminating psychometrics' place in psychological inquiry, but it will be necessary for psychometrics to pursue and maintain interfaces with the other brain sciences to remain relevant to the study of cognition.[50]

6. Causal Inference in Between-subject and Within-subject Contexts

Recall the problem posed by local homogeneity (and the lack thereof). We have local homogeneity just when a model of variability, such as a 1-factor reflective model, fits the population and each individual within that population. Local homogeneity is an assumption that is independent of the model, which is one of the lessons we get from Molenaar's work on ergodicity and dynamic factor models. A consequence of the independence of local homogeneity and the fact that, in general, it is not a feature of measurement models in psychometrics creates an epistemological lacuna between what we can reasonably infer about variability within the population (between subjects) and what we can reasonably infer about variability in the

[50] One potentially promising point of interface is in the field of cognitive IRT modeling. Also, recent work in fMRI imaging has sought to "map" psychometric abilities.

individuals who are members of that population (within subjects). In the context of a discussion of general intelligence, the independence of local homogeneity entails that we cannot, without independent justification, infer that the presence of a *g* in a population-level analysis entails that there will be a *g* when we examine the structure of intraindividual variability; in other words, evidence for general intelligence that comes from population level analyses cannot, alone, settle the question of whether individuals can be said to have general intelligence. Since the *g* factor model is the starting point for the theory of general intelligence, this problem threatens to undermine the foundation of the theory.

6.1 Causal Inference in Between-subject Contexts

Suppose we have a matrix **R** of positive correlations between performance on various tests or items for a population of test takers. **R** exhibits a positive manifold and so we fit a factor model **M** to **R**. Suppose further that **M** admits of a dominant latent factor ξ. Assume that **M** is well confirmed against a set of novel correlation matrices **R′** such that we have high confidence that **M** adequately fits the structure of variability in ability in the population at large.[51] **M** is typically interpreted to be a common cause model with differences in ξ as the common cause of variability in test performance; ξ screens off the correlations between tests (in psychometrics, this feature of the model is called 'local independence'). Interpreted in this way, **M** tells us that differences in ξ in the population cause differences in test performance in the population. What **M** does not tell us is that for an individual S in the population, S's level of ξ causes S's particular item responses. With **M** in hand, we can rank individuals with respect to their level on ξ, but assigning an individual a position on ξ is uninformative about the etiology of that individual's item responses or test performance. This is unsurprising, especially when we consider that **M** is simply a

[51] The arguments presented here will be general, though for the purposes of illustration I will use psychometric abilities and general intelligence as examples of latent traits.

temporal snapshot of population-level differences. An investigation of how particular item responses come about would require that we take the individual, not the population, as the unit of analysis. The between-subject model is simply the wrong tool if what we want to figure out is whether ξ (or, more precisely, an individual's position on ξ) plays some causal role in an individual's test performance.

The reason for talking about the causal efficacy of a level of ξ rather than ξ itself is that the latent dimension ξ itself, i.e., the thing identified in standard factor analysis, is an abstract entity like length or sex. Following the analogy, it is not the dimension that has causal powers in the individual, it is *occupying a position on the dimension* that has causal powers; what causes me to register for the Selective Service is, in part, the fact that I am male—the fact that I occupy a certain value on the dimension 'sex', not sex itself. Sex is simply a dimension along which people differ (dichotomously). In identifying the latent dimension ξ, we have not *ipso facto* identified a cause. We have, in the best-case scenario, identified a dimension along which people vary, and the individual exemplifications of that dimension are the candidates for causal powers. Whether that dimension of variability is of interest is a sociological question. Sex turns out to be a dimension of variability that is of profound importance, both biologically and socially, and hence it generates great interest. Intellectual ability, too, seems to be of great biological and social importance if the apologists of intelligence research are correct. The importance extends beyond the social realm. True, we have a culturally ingrained fascination with IQ scores as evidenced (and possibly constructed) by the attention of popular and scientific media. And, for better or worse, we, arguably, live in a society that imparts a high market value on the very abilities intelligence tests are claimed to measure. It is also likely that the ability to plan and reason abstractly was evolutionarily advantageous, hence its biological importance.

The fact that ξ is causally inert does not mean that it does not have any explanatory role. The fact that we can identify ξ tells us that there are differences between people, i.e., that there is

variability along a single dimension. Borrowing an analogy from Bartholomew (2004), consider Chile and the geographic distribution of its cities. Any city or point in a city can be nearly perfectly located given its 3-dimensional coordinates. But what we find is that there is a dominant dimension along which the cities vary, namely latitude; in terms of information content about a city's location, locating it along this dimension is very informative. While the dimension latitude itself does not have any causal efficacy, variability *in* latitude, a population-level phenomenon, may be causal. Though it is beyond the scope of the present discussion, there are *prima facie* reasons to think that population-level properties such as variability, means, and frequencies can be causal. For example, without variability, there is no natural selection. Income inequality (and hence variance) in a population affects the health of the members of that population (Glymour, 2003; Kawachi *et al.*, 1999). Also being located at a particular latitude within Chile may be causally relevant to other phenomena.

So while ξ, and by implication g, may contain causally relevant information, they are not causes on the account being offered here. Further, since the relevant measurement models are between-subject models, they tell us nothing about the etiology of a particular subject's item responses. What the (unidimensional) measurement model does tell us is that members of the population can be rank ordered with respect to position on the latent dimension along which individuals vary. It remains to be seen how we could detect the causal significance of occupying a certain level on the latent variable. Being male is causal, but is having ability level=j? And how do we detect this causal efficacy within subjects?

6.2 Causal Inference in the Within-subject Model

The within-subject model differs from the between-subject model in design. Rather than our sample being a population of examinees, the sample is a population of *an* examinee over time. Whereas between-subject models report the structure of individual differences in a sample

population at a time, within-subject models report the structure of differences in a single individual *over* time. An approach to investigating the causal efficacy of ξ immediately presents itself: simply examine the covariation of an individual's position on ξ, which we obtained from our population level analysis, and the individual's test performance. But this cannot work. The problem is that one's position on ξ is constant. This does not entail, contrary to what some have said (Holland, 1986; Borsboom, 2005; Borsboom, Mellenbergh, & van Heerden, 2003), that one's position on ξ is not a cause of test performance. Constants can be implicated in causality (*cf.* Dretske, 2004). The issue is epistemological instead of metaphysical. The absence of actual variability in position on ξ makes it difficult to assess whether it is part of the etiology of test performance, but it does not imply that causation is not present. Thus, in testing individual causal claims, we will need some covariation to detect the causal contribution of one's position a latent variable, if it is there. The background of covariation relative to which we will assess the causal efficacy of position on ξ will be counterfactual covariation. For example, the claim

> (1) That Einstein occupies position *i* on general intelligence caused him to answer the item correctly,

will be assessed by evaluating the following counterfactual for a dichotomous item (correct/incorrect):

> (2) If Einstein occupied a lower position on general intelligence he would have answered the item incorrectly.

For the moment, let us set aside the issue of how to determine the truth value of this counterfactual and consider a methodological objection to this approach.

Borsboom (2005) and Borsboom *et al.* (2003) argue that the prospects for a counterfactual account of the within subject causal account of latent variables are not promising. The first objection is that the counterfactual analysis of the causal efficacy of latent variables is, actually, an account of between-subject variability in disguise and, therefore, is uninformative regarding the causal efficacy of the latent variable within subjects. (2) can be reformulated as (3)

> (3) If Einstein had had John's level of general intelligence, [Einstein] would have answered the item incorrectly.

(3) does not seem to express the causal efficacy of general intelligence within Einstein when he answered the item correctly. Rather, according to Borsboom *et al*., (3) expresses a between-subject causal claim that Einstein scored higher than John because Einstein was more intelligent than John. I'll return to this point momentarily. Similarly, Borsboom *et al*. argue that

> (4) If Einstein had had the intelligence of a fruitfly, he would not have been able to answer the item correctly,

expresses a between-subject claim about the differences between the population of humans and the population of fruitflies; (4) tells us nothing about the causal efficacy of Einstein's intelligence in answering the item. No matter how many times we administer Einstein an intelligence test, we will never obtain information regarding the causal efficacy of Einstein's intelligence since the latent variable is inextricably tied to the between-subject measurement model. To assume that a model specifying sources of variability between subjects also contains information regarding how individual scores are produced is to assume local homogeneity. Since the within-subject model tracks the way in which Einstein changes over time, to assume local homogeneity in this case would be tantamount to assuming that the ways in which Einstein varies over time are the same

as the ways in which he differs from others, including fruitflies. Since I am loath to speculate about the cognitive facilities of fruitflies after witnessing their tenacity and cleverness in my kitchen, I will just stick to (3) for the current purposes.

Borsboom *et al.*'s claim that (3) suggests a between-subject causal claim is unclear. The case seems underdescribed. Also, they seem to require that John and counterfactual Einstein be exchangeable in order to test the counterfactual. If this is the locus of their critique, then I agree that there is a problem. There is no experimental access to counterfactual Einstein, and, thus, the claim is untestable. However, there is an alternative approach that seems to offer some promise. Rather than requiring exchangeability between counterfactual and actual entities, the experiments (IRB notwithstanding) may be performable by requiring exchangeability between two actual entities and subjecting one to intellectual manipulation.

Psillos (2004), responding to Holland (1986) and Rubin (1978), advocates such an approach. He offers an experimentally oriented way of testing counterfactual claims such as (2). I will explain his approach and assess whether it offers insight into the causal efficacy of between-subject latent traits. For the sake of continuity, I will tailor Psillos' examples to fit in the psychometric context; this will reveal certain difficulties that arise in the case of determining the causal efficacy of positions on general intelligence. Let the outcome of an intervention on a subject, u, be Y (score on an intelligence test) and let the outcome given no intervention (or the administration of a placebo) be Y′. Suppose the intervention is the administration of some agent that lowers an individual's score on a general intelligence or some other latent cognitive ability such as verbal intelligence. Y–Y′ gives us the effect of lowering intelligence in u. In an actual experimental context this is impossible since you cannot both give and not give u the stupefying agent. If u is given the stupefying agent, t, then the counterfactual scenario where u is not given t, is epistemically out of reach. That is, if we observe the effect of t on u (Y(t, u)), then the control scenario (Y (c, u)) is in principle unobservable. It is for this reason that Holland and Borsboom

claim that without variability there can be no causation; they've let their epistemology do their metaphysics for them. So the question becomes "how can we determine the individual causal effect of intelligence if the counterfactual scenario is inaccessible?" If Y(c, u) is not observable, then we cannot calculate the difference between the control scenario and the treatment scenario.

Psillos' proposed solution to this inference problem is to show that Y(c, u), given certain assumptions, *is* observable. Suppose u receives t. Y(c, u), therefore, is the counterfactual scenario. We give a different subject, u′, the placebo so that instead of being stuck with the inaccessible Y(c, u), we have the observable Y(c, u′). To justify allowing Y(c, u′) to stand in for (counterfactual) Y(c, u), *unit homogeneity* must hold. Given unit homogeneity, u and u′ are matched with respect to all causally relevant factors, i.e., they form a causally homogeneous reference class H. Unit homogeneity may be expressed as follows:

(5) Y(t, u) = Y(t, u′) and Y(c, u) = Y(c, u′).

The exchangeability of u and u′ enables us to reformulate the counterfactual Y(c, u) as a factual claim. Thus, to assess the individual causal effect of t, we determine

(6) Y(t, u) – Y(c, u′).

The outcome of this assessment (purportedly) tells us not only the individual causal effect of t on u, but also the individual causal effect of t on each element of H. The connection to the case of latent variable analysis is straightforward. H would consist of individuals sharing, among other things, membership in the class of individuals initially occupying a certain position *j* on ξ. This gives us one kind of homogeneity. Note that homogeneity with respect to position on the latent

variable alone is not sufficient for generating H. The class of individuals occupying ξ_j is probably *not* causally homogeneous.

Consider the class J of individuals occupying ξ_j. Each member of J has his ability estimated by his performance on the same battery of tests. However, for any set of items, there are multiple ways to achieve the same score, i.e., scores supervene on item performance. Not everyone in J will have the same score pattern necessarily. Perhaps subject A got items 1, 2, 3, 12, and 16 correct, whereas subject B got items 1, 2, 3, 11, and 17 correct. There's no reason to suspect that the same causal processes lead to this different score patterns especially since the content of the items may be different. That the score patterns differ is *prima facie* evidence that A and B achieved their score on ξ through different causal processes since had they been alike in their causal processes, they would have generated the same item responses *ceteris paribus*. What processes are at play is an empirical matter, but there's little justification for the assumption that the same processes were at play when A and B solved the items. In fact, if the items answered correctly are different enough in content, we may have good grounds for suspecting that there are different causal processes at work in the subjects. This suggests that members of J not only occupy the same position on the latent variable, but that they also are homogeneous with respect to response patterns. But now the severity of the problem becomes evident. For any pattern of responses, two subjects who generate that pattern of responses might generate it by way of different causal processes. This situation suggests a solution: check to see if the processes are the same. In such a study we would first have to establish the appropriate level of description. At a high level, we might ask subjects what heuristics they used in solving the items. At a low level, we might monitor neural conductivity or speed of processing. However, we cannot tell whether J is a homogeneous reference class until we have an understanding of ξ_j itself. But if we have an understanding of ξ_j, then, presumably, we know whether it is causally efficacious.

Furthermore, without an antecedent understanding of the processes or properties characteristic of a particular position on a latent variable, an experimental outcome is indeterminate between the causal effect being the result of occupying a different position on the latent variable due to t or the causal effect being the result of t affecting some other causal factor relevant to test performance. Suppose we successfully constructed H and we have two exchangeable individuals so that we can carry out the experiment required to assess the truth of the counterfactual claim about u and we get a non-negligible, statistically significant difference between scores from u and u′. Problem: where in the causal chain between indicators and the latent cause does t operate? The treatment may diminish scores by affecting the ability, it may diminish scores by preempting the causal powers of the ability, e.g., by intervening in the causal chain that runs between indicators and the ability. The treatment may even affect scores by affecting skills that are specific to the particular test. In other words, the treatment may, for each test, affect the portion of the score that is not caused by the ability. Since no indicator is a perfect measure of an ability, there will always be this latter possibility. In cases where the test specific variance is small, but the effect of the treatment is high, we can infer that the treatment did not affect the subject's test specific skill only if we presuppose that the test specific variance that describes the item also applies to the individual. But since test specific variance is a *population* level statistic, it is not clear that the magnitudes that describe the population apply to the individual. Therefore, it seems that a interpretation of a test along the lines Psillos suggests would be underdetermined at least three ways: between affecting the ability, intervening in the causal chain that runs from the ability to item response, and affecting the test specific skill.

Now I will argue that a causally homogeneous reference class with respect to Y, where Y is a certain test score, need not be identical to the class of individuals who are homogeneous with respect to a score on ξ. Suppose we have a causally homogeneous reference class K with respect to some observed score Y. That is, everyone in K has the same test score. Furthermore, by

hypothesis, everyone in K achieved Y by the same kind of process at some level of description. Note that constructing K is no small task, for it will entail identifying the causally relevant processes. Moreover, if the results are to be statistically significant, the initial sample of people with score Y will have to be very large to ensure that the subclass K is representative. There will be people who achieved Y but who are not in K. Some people will get the score Y through different means. Recall that position on ξ is estimated by test score; hence, a causally homogeneous reference class with respect to Y is not identical to the class of individuals who are homogeneous with respect to their score on ξ, since those who achieved Y by some other process (and hence a particular estimated score on ξ) will not be in K. This is just another way of saying that a given level of ξ is multiply realizable with respect to the heuristic, neural, or cognitive processes underlying test behavior. But note that the outcome of the experiment will not establish the causal efficacy of the level of latent trait; rather, it will establish the causal efficacy of certain cognitive processes, namely the ones according to which we partitioned K. The latent trait drops out of the picture.

It might be objected that I've attended to the wrong experimental design in Psillos' discussion. After all, we are searching for the causes of item responses within individuals, and what I've described so far is aimed at quantifying changes *between* individuals. First I will argue that the two designs are not importantly different, given the strict conditions that must be met in order to assign a truth value to a counterfactual. I will argue for this claim by showing that when these conditions are met in the context of making counterfactual claims about levels of ability, they imply local homogeneity of the between-subject measurement model; i.e., local homogeneity is necessary for evaluating the truth value of counterfactual claims. Since local homogeneity is not satisfied in the case of general intelligence, counterfactual claims about general intelligence will be indeterminate with respect to their truth values.

Psillos' other approach, at first glance, seems more in line with the within-subject causal account of latent traits. In this case the population is made up of temporal stages of u. The analysis is longitudinal, or time-series. Exchangeability as it figures in this case is the ability to substitute u at a time with u at a later time. The conditions that license exchangeability in this case are *temporal stability* (u's scores are reliable, *ceteris paribus*) and *causal transience* (the treatment does not affect future replications; in other words, treatment effects are ephemeral). Given temporal stability and causal transience, all relevant time slices of u are matched with respect to all causally relevant factors, thereby forming a causally homogeneous reference class H*. With these assumptions in place, we may assert

(7) Y(t, u (earlier)) = Y(t, u (later)) and Y(c, u (earlier)) = Y(c, u (later))

The exchangeability of u (earlier) and u (later) enables us to reformulate the counterfactual Y(c, u) as a factual claim. Thus to assess the individual causal effect of t, we assess

(8) Y(t, u (earlier/later)) – Y(c, u (later/earlier)).

The objections to the application of the "between-subject" variant of Psillos' account to the psychometric case apply here as well. Showing the empirical equivalence of the designs would be sufficient to make this point. I will do this by arguing from the exchangeability of the subjects. Consider a hypothetical hybrid design. Suppose that u(later) is exchangeable with u(earlier). Thus, temporal stability and causal transience are satisfied. Select some individual u′ such that u′ is unit homogeneous with respect to u(earlier). We let the experiment run and if we've adequately matched our three nominally distinct subjects, we should get identical effect magnitudes (within a tolerable margin of error). But more importantly for the current discussion u(earlier), u(later), and

u′ form a causally homogeneous reference class. For, if u′ is exchangeable with u(earlier), and u(earlier) is exchangeable with u(later), then u′ is exchangeable with u(later); all units are mutually exchangeable. I've appealed to two assumptions in this reasoning: the transitivity of exchangeability and the symmetry of exchangeability, that is

(9) For all x, y, z, if x is exchangeable with y and y is exchangeable with z, then x is exchangeable with z, and

(10) For all x, y, x is exchangeable with y if and only if y is exchangeable with x.

(9) is plausible because the causally relevant features are matched. That is, exchangeability is nothing more than identity with respect to causally relevant features, hence the motivation for (5). I take it as uncontroversial that identity is transitive (within possible worlds). (10) is defensible on the same grounds. If exchangeability is identity with respect to causally relevant features, then exchangeability is symmetric if and only if identity is, which it is. In terms of Psillos' schema, we get the following:

(11) Y(t, u (earlier)) – Y(c, u (later)) = Y(t, u) – Y(c, u′).

Therefore the two designs, though methodologically distinct, are empirically equivalent.

A lesson to be learned from this discussion is that order to assess whether psychometric abilities are causal, we have to have some grasp on what they are, independent from their statistical characteristics. Statistical characteristics tell us nothing about constitution or mechanisms underlying ability. The problem is that the ability is more than just how it is realized in any one causally homogeneous reference class. Further difficulties arise when we consider the

fact that we have investigated the causal efficacy of only one level of the ability. Hence, there is not only the problem of generality at a single level of ability, but also the problem of generalizing across levels of ability. The ways in which the ability is differentially realized at a given level may be quite different from how it is realized across levels. Development is also an important complication to consider. For example Esposito *et al.* (1999) conducted PET studies that revealed that age was relevant to what neural networks were activated in a particular task. Additionally, We must be sensitive to the fact that these levels may be only loosely connected with neurophysiological differences since positions on ability are ordinal rankings. One's level on some ability is fixed relative to others on the basis of performance. Of course there is a reason why a particular subject consistently scores at a certain level and this will be partly explained by in terms of cognitive mechanisms or neurological facts since they will explain why the subject got a particular raw score. But also among the *explanans* is the fact that the rest of the population performed in a certain way. Consider the possibility that the normative sample is not representative in that there are actually more people scoring at the high end than the normative sample indicates. Test developers notice this and renorm the test. After renorming, a person's raw score that before was, say, average would correspond to a lower level of the ability. The raw score remains stable, as do the cognitive processes that resulted in that raw score, but the norm-referenced score changes in light of the revised normative data. Thus, it seems that the actual *explanandum* for brain science is *not* the level of ability. It is, rather, the raw score. Explaining why someone occupies a particular level may be of interest to pathologists who want to explain abnormality (a norm referenced concept), but that explanation will involve not only explaining the intrinsic characteristics of the pathology, but also how those characteristics differ from the intrinsic characteristics of nonpathological cases.

6.3 Reflections on Unit Homogeneity, Temporal Stability, and Local Homogeneity

Let us set aside the above concerns for a moment and consider the conditions that must be met for causal inferences to be warranted on this account. Unit homogeneity, temporal stability, and causal transience are the conditions that must be met by subjects in order to construct a causally homogeneous reference class and, thus, to determine the truth value of a counterfactual claim such as (2). We want to determine the truth value of (2) so that we can justify or determine the truth value of (1). Unit homogeneity is the condition that the individuals in an experiment are similar enough to respond to t in the same way. Modeled statistically, the subjects would be governed by the *same dynamical laws*; the within-subject variability of u and u′ after receiving t would be the same. As Psillos writes, "[unit homogeneity] simply means that that there is *a* causal law connecting the treatment and its characteristic effect which holds for *all* homogeneous units and hence is independent of the actual unit chosen…to test it," (2004, p. 305).[52] In the repeated measures case, we required that u(early) and u(later) satisfy the temporal stability requirement (in addition to causal transience). Temporal stability, psychometrically speaking, is reliability. That is, it is consistency of responses. Every time u gets t, u responds the same way. This will have to be relaxed somewhat since measurement error and the probabilistic nature of item response models introduces a stochastic element into responses. This minor qualification aside, we see that temporal stability requires that any developmental characteristics of u must be causally irrelevant to u's response to t. Therefore, u, with respect to, c and t, is a nondevelopmental system and likewise for any other homogeneous unit. Furthermore, temporal stability of response entails u's responses to c and t are invariant with respect to time.

Interestingly, these implications of unit homogeneity and temporal stability are the very two conditions that Molenaar says are necessary and jointly sufficient for local homogeneity. If u and u′ are unit homogeneous, then they satisfy Molenaar's condition of *homogeneity of the*

[52] Italics added.

ensemble. Homogeneity of the ensemble is satisfied when the units in the ensemble obey the same dynamical laws. If u(early) and u(later) exhibit temporal stability, then they are invariant with respect to time, or, in Molenaar's terminology, the processes that give rise to the response exhibit *time-invariant sequential dependencies*. Thus, our causally homogeneous reference class of units exhibits local homogeneity. This is not surprising when we consider that our ensemble is made up of virtually identical units. However, recall that I argued earlier that *g* factor models are not locally homogeneous. Does this present a conceptual obstacle to testing (1) or (2)? First note that if we were to give a battery of tests to the units of the ensemble, no psychometric ability would be derivable for the same reason that no individual will be described by a *g* factor model: there is not the requisite variability. Any variability that arises is likely to be error variance and not variance due to a common factor. This would be problematic if we were trying to test the causal efficacy of *g*; however, I've argued that *g* is noncausal. It is plausibly construed as an abstract dimension such as latitude or sex. What is being tested is the causal efficacy of one's position on the ability (assuming a normal distribution and representative norm sample) or raw score with respect to test performance. We run afoul of the lack of local homogeneity only if we assume that *g* has a causal force since it is a variable defined *over* individuals, not *within* individuals. Conversely, we should not look at the between-subject model and conclude that it tells us anything about the nature of the psychological processes at play in individuals. Nevertheless, the between-subject models are not completely detached from the elements of the ensemble. After all, without the ensemble, there would be no between-subject model.

The between-subject model tells us is that there is a dimension along which people vary. That dimension is abstract and noncausal. To identify the salient processes, we should find tasks that load highly on that dimension, for they will be the ones that maximally discriminate along that dimension. The between-subject model gives us a kind of sorting machine. When there are differences in exemplifications of the dimension (i.e., different factor scores) we do the appropriate brain science to identify the characteristic cognitive functions associated with that

exemplification. This tells us the intraindividual processes at work in solving problems contained in psychometric tests. With this information in the background, can then ask the question "Why do subjects at one level perform differently than subjects at another level?" Having identified the intraindividual processes, we can then perform interindividual comparisons with respect to the intraindividual processes. Therefore, the population-level analysis, while it cannot be an account of what happens in the individual, nevertheless provides constraints on and guidance in studies into the biological basis of individual differences by delivering the initial *datum*, the dominant dimension on which individuals differ.

7. Validity Revisited

The fact that a model is (or is not) locally homogeneous is not sufficient on its own to settle the question of realism regarding some attribute, nor the question of whether the attribute being measured in a population is the same attribute being measured in the members of that population. Sex may be "real" even if it is not causal and even if it cannot be modeled in the individual. Sex is real in the sense that it is a dimension along which people vary in virtue of occupying different levels on the trait. Sex is not causally efficacious, but occupying one of its position (e.g., being female) is. Nevertheless, local homogeneity (or the lack thereof) can, as it did above, point research in certain directions that may have consequences test validity. At this point I turn to a discussion of where things stand with respect to general intelligence and validity.

According to Borsboom (2005; Borsboom *et al*. 2004, p. 1061), a test is valid for measuring an attribute if and only if (a) the attribute exists (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure (p. 1061). Therefore, a test of general intelligence is valid if and only "[i]f general intelligence exists, and the IQ-test items have the property of transmitting variation in general intelligence, [in which case] we can [then] use somebody's responses on these items to assess his position on general intelligence," (Borsboom 2005, p. 146). The structure of between-subject variability may be different from the structure of

within-subject variability, and yet the two analyses may point to the same attribute. But it is an empirical question whether this is the case. It holds for height, but it may not hold for psychological attributes. The above quotation reveals a potential shortcoming in Borsboom's analysis of validity. It applies only to the between-subject model. Administered to an individual, IQ items will transmit no variation in general intelligence, not even after repeated measures (neglecting practice effects). Furthermore, if attributes like general intelligence are noncausal, as I've argued and as Borsboom agrees (personal communication), then it cannot cause variations in the outcomes in of the measurement procedure. Variability in test scores may have some causal effect on the population just as income inequality can have effects on individuals or populations, but I take it that variability alone is not what interests psychometricians or the clinical psychologist who administers IQ tests to children. I can assent to (a) provided it means nothing more than that there is a dominant dimension along which people differ. No reification is necessary since I do not ascribe causal powers to that dimension. This is realism about psychometric attributes, albeit in a form much weaker than that which Gould ascribes to Jensen, Spearman, Burt, and others. A viable conception of validity should address not only between-subject models, but also measurement of individual psychological attributes independent of the individual's position relative to others. Without local homogeneity, there is always the possibility that the attribute identified in between-subject analyses of performance is not the attribute implicated in individual assessment. Recalling Molenaar's quasi-experimental studies, we can say that a test of personality is valid in Borsboom's sense (for the population), but the significance of the study was that the attribute structure failed to apply to any individual. Paradoxically, it seems that Molenaar's tests were valid for the population, but not for the individual subjects. Psychometricians need an account of validity that is responsive to this concern, one that accommodates the between-subject model (such as Borsboom's), but also applies to singular instances of measurement. The latter account cannot be statistical. It must appeal to causal processes in individual at a single measurement occasion.

8. Conclusion

There's valuable methodological advice to be gleaned from this discussion. It turns out that in the course of testing the counterfactual it will be necessary to identify the specific processes responsible for getting a particular test score so that we have a causally homogenous reference class with respect to the score. But once this is done, we can do away with talk about latent abilities in favor of less objectionable entities that we can actually observe and readily manipulate. The quest to establish the causal efficacy of the latent ability has borne fruit, though not the kind we might have suspected (or that the psychometricians may have wanted). Nevertheless there's an important lesson here. Psychometricians cannot expect to go far in understanding cognition or the empirical significance of their theoretical posits unless they coordinate their efforts with other disciplines in the brain sciences. That ideally we would do away with latent variables as explanatory posits in psychology is no threat to the psychometric enterprise. Recall, it is noting that there are individual differences in test scores that led us down this road to begin with. Identifying these differences and evaluating their statistical significance is an essential part of the process of coming to understand how cognition works. We then, naturally, seek to explain why there were these differences. The psychometric results can be the starting point for substantive psychological inquiry.

In the course of this study, I've shown that psychometricians are not necessarily the crude sort of unreflective reifiers of latent variables that Gould, mistakenly, accuses Jensen and Spearman of being. There is latent variable realism in psychometrics and I've argued that it is the most plausible philosophical position for making sense of psychometric practice. But this realism need not be "pathological" (in Michell's sense) or otherwise scientifically illegitimate. Epistemic realism, as embodied in psychometric practice, represents a willingness to listen to evidence and, at worst, unjustified optimism about what fruits psychometric inquiry can yield. Commitment to the existence of general intelligence or extroversion, for example, is inextricably tied up with

confirmatory statistical tools that provide (or undermine) justification for ontological realism. so that ontological commitment is based in evidence. I've considered two accounts of test validity, both of which are realist in character, though to different extents. On pragmatic and conceptual grounds I argued that Borsboom's test-based approach to validity is preferable to Messick's inference based approach. However, I noted that even Borsboom's account leaves somethings to be desired, such as an account of what it means to say a psychological attribute, e.g., general intelligence, exists. Some may think it a flaw in Borsboom account that it is not a comprehensive theory of validity and validation. But Borsboom's account is realist to the core, and a concomitant theory of validation would, in essence, give the conditions under which one would be justified in believing that an unobservable psychological attribute exists. I argue that while such an account is desirable, it may be a bit much to expect of an account of validity that it settle the realism debate.

In this chapter I returned to the topic of validity spelled out some of my concerns regarding Borsboom's account of validity and noted that if we are to be realists about psychological attributes like general intelligence, then our realism must be suitably tempered in light of the problem of local homogeneity. A consequence of this problem is that general intelligence, if real, is not a causally efficacious attribute of individuals; it is a population-level phenomenon. As such, it may have causal powers, but they are not the ones psychologists would likely ascribe to it. Considering work on causation in epidemiological studies, I argued that there is no principled reason to deny that a between-subject phenomenon can be causal. But note that if we conceive of general intelligence and, say, income inequality analogously, general intelligence ceases to be a feature of individuals. It simply makes no sense to say that S achieved score Y because of his general intelligence. The search for a person-specific quality causally responsible for test score led me to consider the *g*-score. I applied an approach championed by Psillos and Rubin for testing counterfactual claims implied by causal claims. Though this approach shows promise in rendering the truth values of counterfactuals determinable, it only uncovered practical obstacles in its application to the *g*-score. What is lacking, ultimately, is an account of the

processes underlying the *g*-score. Without a specification of its biological basis, the *g*-score is lean on explanatory resources. One interesting result of this investigation is that the conditions for testing causal claims about the *g*-score are met just when the conditions for local homogeneity hold. Consequently, satisfying local homogeneity does not bring with it an account of the intraindividual processes that bring about item responses.

I have also argued that realism, despite what critics of the realism debates would say, matters and bears on psychological theory. The theory of general intelligence is essentially bound up with certain statistical tools. These tools give us the statistical models that make us think there is any such thing as general intelligence at all. However, the models that they deliver are undetermined by the evidence. For a given set of data, there are many different models that fit it. These models when interpreted from the realist point of view say different things about cognitive abilities, however. In Chapter 4 I argued that one way in which realism matters is that it guides model adjudication. A realist about measurement models in psychometrics and Jensen's theory of general intelligence can rule out the bifactor model of intelligence in favor of the higher order factor model. Psychometric and pragmatic reasons for preferring the higher order model were also discussed.

Studies in scientific realism are not as popular now as they were fifteen or twenty years ago, and if you ask someone intimately familiar with the debate why this is, the person is likely to respond (I've found) by saying that philosophers simply realized that it doesn't matter if one is a realist or not. The debate has no bearing on science, the person will continue. I have two things to say in response to this. First, while I see it as a virtue of this dissertation that it seeks to engage with the relevant science and effect change therein, I do not, in general, see why we should accept a criterion of significance in the philosophy of science (or philosophy generally) according to which tangible (rather than conceptual) consequences are paramount. Second, I hope that this dissertation has shown that there is still interesting work to be done in realism studies, especially in the less mature sciences such as psychology.

BIBLIOGRAPHY

Aizawa, Kenneth, and Carl Gillett (unpublished manuscript), "Multiple Realization and Methodology in the Neurological and Psychological Sciences".

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1985), *Standards for Educational and Psychological Testing*. Washington DC: American Psychological Association.

——— (1999), *Standards for Educational and Psychological Testing*. 2nd ed. Washington DC: American Psychological Association.

Anderson, Mike (1992), *Intelligence and Development*. Cambridge: Blackwell.

Azzelini, A. (1996), *Statistical Inference Based on the Likelihood*. London: Chapman and Hall.

Baldwin, A.L. (1946), "The Study of Individual Personality by Means of the Intraindividual Correlation", *Journal of Personality* 14:151-168.

Bartholomew, David (2004), *Measuring Intelligence: Facts and Fallacies*. New York: Cambridge.

Beaujean, A. Alexander (2002), "SASP Interviews: Arthur R. Jensen", in, *SASP News*, 8–19.

Bechtel, William (2007), "Reducing Psychology while Maintaining its Autonomy", in M. Schouten and H. Looren de Jong (eds.), *Matter of the Mind: Philosophical Essays on Psychology, Neuroscience, and Reduction*, Oxford: Basil Blackwell.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell (1975), "Sex Bias in Graduate Admissions: Data from Berkeley", *Science* 187:398–404.

Bickley, P. G., T. Z. Keith, and L. M. Wolfe (1995), "The Three-stratum Theory of Cognitive Abilities: Test of the Structure of Intelligence Across the Life Span", *Intelligence* 20:309–328.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*. New York: John Wiley.

Boring, Edwin G. (1923), "Intelligence as the Tests Test It", *New Republic*, June 6, 35-37.

Borsboom, Denny (2005), *Measuring the Mind: Contemporary Issues in Psychometrics*. Cambridge: Cambridge University Press.

——— (2006), "Attack of the Psychometricians", *Psychometrika* 71 (3):425-440.

Borsboom, Denny, and Conor V. Dolan (2006), "Why g Is Not an Adaptation: A Comment on Kanazawa (2004)", *Psychological Review* 113 (2):433-437.

Borsboom, Denny, Gideon Mellenberg, and Jaap van Heerden (2003), "Validity and Truth", in H. Yanai, A. Okada, K. Shigemasu, Y. Kano and J. J. Meulman. (ed.), *New Developments in Psychometrics. Proceedings of the International Meeting of the Psychometric Society 2001*, Tokyo: Springer.

——— (2003), "The Theoretical Status of Latent Variables", *Psychological Review* 110 (2):203-219.

Borsboom, Denny, and Gideon Mellenbergh (2004), "Why Psychometrics is not Pathological: A Comment on Mitchell", *Theory & Psychology* 14 (1):105-120.

Borsboom, Denny, Gideon Mellenbergh, and Jaap van Heerden (2004), "The Concept of Validity", *Psychological Review* 111 (4):1061-1071.

Boyd, Richard, Philip Gasper, and J.D. Trout, eds. (1991), *The Philosophy of Science*. Cambridge: MIT Press.

Bridgman, Percy (1927), *The Logic of Modern Physics*. New York: Macmillan.

——— (1991), "The Operational Character of Scientific Concepts", in Richard Boyd, Philip Gasper, J.D. Trout (ed.), *The Philosophy of Science*, Cambridge: MIT Press.

Byrne, B. M., and M. S. Sunita (2006), "The MACS Approach to Testing Multigroup Invariance of a Second-order Structure", *Structural Equation Modeling* 13 (2):287–321.

Carroll, J. B. (1993), *Human Cognitive Abilities: A Survey of Factor Analytic Studies*. New York: Cambridge University Press.

——— (1995), "Reflections on Stephen Jay Gould's The Mismeasure of Man (1981)", *Intelligence* 21:121-134.

——— (1997), "Theoretical and Technical Issues in Identifying a Factor of General Intelligence", in B. Devlin, S.E. Fienberg, D.P. Resnick and K. Roeder (eds.), *Intelligence, Genes, & Success. Scientists Respond to The Bell Curve*, New York: Copernicus.

Carson, John (2007), *The Measure of Merit: Talents, Intelligence, and Inequality in the French and American Republics, 1750-1940*. Princeton: Princeton University Press.

Cartwright, N. (1983), *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Cattell, R. B. (1946), *Description and Measurement of Personality*. New York: World Book Company.

——— (1971), *Abilities: Their Structure, Growth, and Action*. Boston: Houghton Mifflin.

Cervone, D. (2004), "The Architecture of Personality", *Psychological Review* 111:183-204.

Chabris, Christopher F. (2006), "Cognitive and Neurobiological Mechanisms of the Law of General Intelligence", in M. J. Roberts (ed.), *Integrating the Mind*, Hove, UK: Psychology Press.

Chen, F. F. , S. G. West, and K. H. Sousa (2006), "A Comparison of Bifactor and Second-Order Models of Quality of Life", *Multivariate Behavioral Research* 21 (2):189–225.

Colom, R. , I. Rebollo, A. Palacios, M. Juan-Espinosa, and P. C. Kyllonen (2004), "Working Memory is (almost) Perfectly Predicted by g", *Intelligence* 32:277–296.

Conway, A. R. A., N. Cowan, M. F. Bunting, D. J. Therriault, and S. R. B. Monkoff (2002), "A Latent Variable Analysis of Working Memory Capacity, Short-term Memory Capacity, Processing Speed, and General Fluid Intelligence", *Intelligence* 30 (163–183).

Costa, P.T., and R.R. McCrae (1992), *Revised NEO Personality Inventory (NEO PI-R)*. Odessa, FL: Psychological Assessment Resources.

Crocker, Linda, and James Algina (1986), *Introduction to Classical and Modern Test Theory*. New York: Harcourt Brace Jovanovich College Publishers.

Cronbach, L.J. (1949), *Essentials of Psychological Testing*. New York: Harper and Brothers.

Cronbach, L.J., and P.E. Meehl (1955), "Construct Validity in Psychological Tests", *Psychological Bulletin* (52):281-302.

de Koning, E., K. Sijtsma, and J. H. M. Hamers (2003), "Construction and Validation for Test of Inductive Reasoning", *European Journal of Psychological Measurement* 19 (1):24–39.

Deary, I. J. (2001), *Intelligence: A Very Short Introduction*. Oxford: Oxford University Press.

——— (2002), "g and Cognitive Elements of Information Processing: An Agnostic View", in R. J. Sternberg and E. L. Grigorenko (eds.), *The General Factor of Intelligence: How General is it?*, Mahwah: Erlbaum, 151-182.

Deary, I. J. (2000), *Looking Down on Human Intelligence*. Edited by Mackintosh et al., *Oxford Psychology Series*. Oxford: Oxford University Press.

Deary, I. J., and Claudia Pagliari (1991), "The Strength of g at Different Levels of Ability: Have Detterman and Daniel Rediscovered Spearman's "Law of Diminishing Returns"?" *Intelligence* 15:247-250.

Detterman, D. K. (1991), "Reply to Deary and Pagliari: Is g Intelligence or Stupidity?" *Intelligence* 15:251-255.

Detterman, D. K., and Mark H. Daniel (1989), "Correlations of Mental Tests with Each Other and with Cognitive Variables are Highest for Low IQ Groups", *Intelligence* 13:349-359.

Dickens, William T., and James R. Flynn (2001), "Heritability Estimates Versus Large Environment Effects: The IQ Paradox Solved", *Psychological Review* 108 (2):346-369.

——— (2002), "The IQ Paradox is Still Resolved: Reply to Loehlin (2002) and Rowe and Rodgers (2002)", *Psychological Review* 109 (4):764-771.

Dolan, C.V., R. Colom, F.J. Abad, J. Wicherts, D.J. Hessen, and S. van der Sluis (2006), "Multi-group covariance and mean structure modeling of the relationship between WAIS-III common factors and gender and educaation attainment in Spain", *Intelligence* 34 (2):193–210.

Dretske, Fred (2004), "Psychological vs. Biological Explanations of Behavior", *Behavior and Philosophy* 32:167–177.

Edwards, J. R., and R. P. Bagozzi (2000), "On the Nature and Direction of Relationships between Constructs and Measures", *Psychological Methods* 5:155-174.

Ellis, J. L., and A. L. van den Wollenberg (1993), "Local Homogeneity in Latent Trait Models. A Characterization of the Homogeneous Monotone IRT Model", *Psychometrika* 58 (3):417-429.

Esposito, G., B. S. Kirkby, J. D. Van Horn, T. M. Ellmore, and K. F. Berman (1999), "Context-dependent, Neural System-specific Neurophysiological Concomitants of Ageing: Mapping PET Correlates During Cognitive Activation", *Brain* 122:963–979.

Eysenck, H.J., and S.B.G Eysenck (1991), *Manual for the EPQ-R*. Sevenoaks: Hodder and Stoughton.

Fagan, M. B. (2007), *Objectivity in Practice: Integrative Social Epistemology of Scientific Inquiry.* Dissertation. Department of History and Philosophy of Science, Bloomington: Indiana University.

Farquhar, I. E. (1964), *Ergodic Theory in Statistical Mechanics*, *Monographs in Statistical Physics and Thermodynamics*. New York: Interscience Publishers.

Ferrando, P. J. (2002), "Theoretical and Empirical Comparisons between Two Models for Continuous Item Responses", *Multivariate Behavioral Research* 37 (4):521–542.

Flynn, James R. (1984), "The Mean IQ of Americans: Massive Gains", *Psychological Bulletin* 95:29-51.

——— (1987), "Massive IQ Gains in 14 Nations: What IQ Tests Really Measure", *Psychological Bulletin* 101:171-191.

——— (1999), "Searching for Justice: The Discovery of IQ Gains Over Time", *American Psychologist*:5-20.

Gardner, Howard (1993), *Frames of Mind*: Basic Books.

Gignac, G. E. (2005), "Revisiting the Factor Structure of the WAIS-R: Insights Through Nested Factor Modeling", *Assessment* 12:320–329.

——— (2005), "The Dimensionality of the WAIS-III: Extensions Using the Bi-factor Model", in.

——— (2006), "A Confirmatory Examination of the Factor Structure of the Multidimensional Aptitude Battery", *Educational and Psychological Measurement* 66 (1):136–145.

Glymour, Bruce (2003), "On the Metaphysics of Probabilistic Causation: Lessons from Social Epistemology", *Philosophy of Science* 70:1413–1423.

Glymour, Clark (1998), "What Went Wrong? Reflections on Science by Observation and the Bell Curve", *Philosophy of Science* 65 (1):1-32.

——— (2001), *The Mind's Arrows*. Cambridge: MIT Press.

Gottfredson, L. S. (2002), "g: Highly General and Highly Practical", in R. J. Sternberg and E. L. Grigorenko (eds.), *The General Factor of Intelligence: How General is It?*, Mahwah: Erlbaum.

Gould, Stephen Jay (1996), *The Mismeasure of Man*. New York: W.W.Norton & Company. Original edition, 1981.

Gregory, Robert J. (1999), *Foundations of Intellectual Assessment: The WAIS-III and Other Tests in Clinical Practice*. Boston: Allyn and Bacon.

——— (2004), *Psychological Testing: History, Principles, and Applications*. 4th ed. ed. Boston: Pearson Education Group, Inc.

Guilford, J. P. (1967), *The Nature of Human Intelligence*. New York: McGraw-Hill.

Gustafsson, J. -E., and G. Balke (1993), "General and Specific Abilities as Predictors of School Achievement", *Multivariate Behavioral Research* 28:407–434.

Hacking, Ian (1983), *Representing and Intervening*. Cambridge: Cambridge University Press.

——— (1986), "Making Up People", in Morton Sosna Thomas C. Heller, and David E. Wellbery (ed.), *Reconstructing Individualism: Autonomy, Individuality, and the Self in Western Thought*, Stanford: Stanford University Press.

Hamaker, E. L. (2004), *Time Series Analysis and the Individual as the Unit of Psychological Research*. Dissertation. Department of Psychological Methods Amsterdam: University of Amsterdam.

Hambleton, Ronald K., H. Swaminathan, and Jane H. Rogers (1991), *Fundamentals of Item Response Theory*. Edited by Richard Jaeger, *Measurement Methods for the Social Sciences*. Newberry Park: Sage.

Herrnstein, Richard, and Charles Murray (1994), *The Bell Curve: Intelligence and Class Structure in America*. New York: Free Press.

Holizinger, K. J., and F. Swineford (1937), "The Bi-factor Method", *Psychometrika* 2:41–54.

Holland, P.E. (1986), "Statistics and Causal Inference", *Journal of the American Statistical Association* 81:945-959.

Hood, S. Brian (forthcoming), "A Comment On Borsboom's Typology of Measurement Theoretic Variables and Michell's Assessment of Psychometrics as "Pathological Science", *Measurement: Interdisciplinary Research and Perspectives*.

Horwich, Paul (2004), *From a Deflationary Point of View*. Oxford: Clarendon Press.

Jensen, Arthur (1980), *Bias in Mental Testing*. New York: Free Press.

——— (1982), "The Debunking of Scientific Fossils and Straw Persons", *Contemporary Education Review* 1 (2):121-135.

——— (1998), *The g Factor*. Westport: Praeger.

——— (2002), "Psychometric g: Definition and Substantiation", in R. J. Sternberg and E. L. Grigorenko (eds.), *The General Factor of Intelligence: How General Is It?*, Mahwah: Erlbaum, 39-54.

Jensen, Arthur, and Li-Jen Weng (1994), "What Is a Good g?" *Intelligence* 18 (1):231-258.

Johnson, W. , T. J. Bouchard, R. F. Krueger, M. McGue, and I. I. Gottesman (2004), "Just One g: Consistent Results from Three Test Batteries", *Intelligence* 32:95–107.

Kanazawa, S. (2004), "General Intelligence as a Domain-specific Adaptation", *Psychological Review* 111:512-523.

Kane, Michael T. (2006), "Validation", in Robert Brennan (ed.), *Educational Measurement*, Washington DC: American Council on Education and National Council on Measurement in Education.

Kavale, K. A., and S. R. Forness (1995), *The Nature of Learning Disabilities: Critical Elements of Diagnosis and Classification*. Mahwah, NJ: Erlbaum.

Kawachi, Ichiro, Bruce Kennedy, and Wilkinson (1999), *The Society and Population Health Reader; Vol. 1, Income Inequality and Health*. New York: The New Press.

Kelley, T. L. (1927), *Interpretation of Educational Measurements*. New York: Macmillan.

Kim, Jaegwon (2006), *Philosophy of Mind*. Cambridge: Westview Press.

Kitcher, Philip (2001), *Science, Truth, and Democracy*. Oxford: Oxford University Press.

Kline, Paul (1976), *Psychological Testing*. New York: Crane Russak.

——— (1993), *The Handbook of Psychological Testing*. New York: Routledge.

——— (1998), *The New Psychometrics*. New York: Routledge.

Kyllonen, P. C. (1996), "Is Working Memory Capacity Spearman's g?" in I. Dennis and P. Tapsfield (eds.), *Human abilities: Their nature and measurement*, Mahwah, NJ: Lawrence Erlbaum, 49–76.

Kyllonen, P. C., and R. E. Christal (1990), "Reasoning Ability is (little more than) Working Memory Capacity?" *Intelligence* 14:389–433.

Lakatos, Imre (1970), "Falsification and the Methodology of Scientific Research Programmes", in I. Lakatos and A. Musgrave (eds.), *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, 91-196.

Lawley, and Maxwell (1971), *Factor Analysis as a Statistical Method*. London: Butterworth.

Leplin, Jarrett (1984), *Scientific Realism*. Berkeley, CA: University of California Press.

——— (1986), "Methodological Realism and Scientific Rationality", *Philosophy of Science* 53:31-51.

——— (1997), *A Novel Defense of Scientific Realism*. New York: Oxford University Press.

——— (2004), "Predictive Success can Warrant Belief in Unobservables", in Christopher Hitchcock (ed.), *Contemporary Debates in Philosophy of Science*, Oxford: Blackwell Publishing.

——— (2005), "Scientific Realism", in Sahorta Sarkar (ed.), *The Philosophy of Science, An Encyclopedia*, London: Routledge, 686-698.

Letho, J. E., Juujärvi, L. Kooistra, and L. Pulkkinen (2003), "Dimensions of executive functioning: Evidence from children", *British Journal of Developmental Psychology* 21:59–80.

Lloyd, E. A. (2005), *The Case of the Female Orgasm: Bias in the Science of Evolution*. Cambridge: Harvard University Press.

Longino, H. E. (1990), *Science as Social Knowledge*. Princeton: Princeton University Press.

——— (2002), *The Fate of Knowledge*. Princeton: Princeton University Press.

Lord, F. M., and M. R. Novick (1968), *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley.

Lovett, B. J., and L. J. Lewandowski (2006), "Gifted Students with Learning Disabilities: Who are They?" *Journal of Learning Disabilities*.

Mackintosh, N. J. (1998), *IQ and Human Intelligence*. New York: Oxford University Press.

Markus, Keith (1998), "Science, Measurement, and Validity: Is Completion of Samuel Messick's Synthesis Possible?" *Social Indicators Research* 45:35-44.

Mellenbergh, G. J. (1989), "Item Bias and Item Response Theory", *International Journal of Educational Research* 13:127–143.

——— (1994), "A Unidimensional Latent Trait Model for Continuous Item Responses", *Multivariate Behavioral Research* 29:223–236.

——— (1994), "General Linear Response Theory", *Psychological Bulletin* 115:300–307.

——— (1996), "Measurement Precision in Test Score and Item Response Models", *Psychological Methods* 1 (3):293-299.

——— (1999), "Measurement Models", in H.J. Ader and G.J. Mellenbergh (ed.), *Research Methodology in the Social, Life, and Behavioral Sciences*, London: Sage.

Meredith, W. (1993), "Measurement Invariance, Factor Analysis, and Factorial Invariance", *Psychometrika* 58:525–543.

Messick, S. (1989), "Meaning and Values in Test Validation: The Science and Ethics of Assessment", *Educational Researcher* 18 (2):5-11.

——— (1989), "Validity", in R.L. Linn (ed.), *Educational Measurement*, Washington DC: American Council on Education and National Council on Measurement in Education, 13-103.

——— (1995), "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performance as Scientific Inquiry into Score Meaning", *American Psychologist* 50 (9):741749.

——— (1998), "Test Validity: A Matter of Consequence", *Social Indicators Research* 45:35-44.

Michell, Joel (1997), "Quantitative Science and the Definition of Measurement in Psychology", *British Journal of Psychology* 88:355-383.

——— (1999), *Measurement in Psychology: A Critical History of a Methodological Concept*. Edited by Quentin Skinner, *Ideas in Context*. Cambridge: Cambridge University Press.

——— (2000), "Normal Science, Pathological Science and Psychometrics", *Theory & Psychology* 10 (5):639-667.

——— (2004), "Item Response Models, Pathological Science and the Shape of Error", *Theory & Psychology* 14 (1):121-129.

——— (forthcoming), "Is Psychometrics Pathological Science?" *Measurement: Interdisciplinary Research and Perspectives*.

Miele, Frank (2002), *Intelligence, Race, and Genetics: Conversations with Arthur Jensen*. Boulder, CO: Westview Press.

Molenaar, P. C. M. (1985), "A Dynamic Factor Model for the Analysis of Multivariate Time Series", *Psychometrika* 50:181-202.

——— (1997), "Time-series Analysis and its Relationship with Longitudinal Analysis", *International Journal of Sports Medicine* 18:232-237.

——— (1999), "Longitudinal Analysis", in H. J. Ader and G. J. Mellenberg (eds.), *Research Methodology in the Social, Behavioural, and Life Sciences*, London: Sage, 143-167.

Molenaar, P. C. M., and A. von Eye (1994), "On the Arbitrary Nature of Latent Variables", in A. and C.C. Clogg von Eye (ed.), *Latent Variables Analysis*, Thousand Oaks: Sage.

Molenaar, P. C. M., H.M. Huizenga, and J.R. Nesselroade (2003), "The Relationship Between the Structure of Inter-individual and Intra-individual Variability: A Theoretical and Empirical Vindication of Developmental Systems Theory", in U.M. Staudinger and U. Lindenberger (ed.), *Understanding Human Development*, Dordrecht: Kluwer.

Neisser, U. (1999), "Two Views About the g Factor: The Great g Mystery", *Contemporary Psychology APA Review of Books* 44 (2):131-133.

———, ed. (1998), *The Rising Curve*. Washington DC: American Psychological Association.

Neisser, U., G. Boodoo, T. J. Bouchard, W. A. Boykin, N. Brody, S. J. Ceci, H. F. Diane, J. C. Loehlin, R. Perloff, R. J. Sternberg, and S. Urbina (1996), "Intelligence: Knowns and Unknowns", *American Psychologist* 51:77–101.

Neisser, U., G. Boodoo, T.J. Bouchard, W.A. Boykin, N. Brody, S.J. Ceci, H.F. Diane, J.C. Loehlin, R. Perloff, R.J. Sternberg, and S. Urbina (1996), "Intelligence: Knowns and Unknowns", *American Psychologist* 51:77-101.

Psillos, Stathis (1999), *Scientific Realism: How Science Tracks Truth*. New York: Routledge.

——— (2004), "A Glimpse of the Secret Connexion: Harmonizing Mechanisms with Counterfactuals", *Perspectives on Science* 12 (3).

Putnam, Hilary (1979), *Mind, Language, and Reality; Philosophical Papers*. Vol. 2. Cambridge: Cambridge University Press.

Raven, J. C. (1938), *Progressive Matrices*. London: H.K. Lewis.

Rindskopf, D., and T. Rose (1988), "Some Theory and Applications of Confirmatory Second-order Factor Analysis", *Multivariate Behavioral Research* 23 (51–67).

Roberts, M. J., ed. (2006), *Integrating the Mind*. Hove, UK: Psychology Press.

Ruelle, David (1991), *Chance and Chaos*. Princeton: Princeton University Press.

Rushton, J. P. (1999), "Secular Gains in IQ Not Related to the g Factor and Inbreeding Depression Unlike Black-White Differences: A Reply to Flynn", *Personality and Individual Difference* 26:381-389.

Sattler, Jerome M. (2001), *Assessment of Children: Cognitive Applications*. San Diego: Jerome M. Sattler, Publisher, Inc.

Schermelleh-Engel, K., H. Moosbrugger, and H. Müller (2003), "Evaluating the Fit of Structural Equation Models: Test of Significance and Descriptive Goodness-of-fit Measures", *Methods of Psychological Research* 8:23–74.

Schmid, J., and J. M. Leiman (1957), "The Development of Hierarchical Factor Solutions", *Psychometrika* 22 (1):53–61.

Sesardic, Neven (2000), "Philosophy of Science That Ignores Science: Race, IQ and Heritability", *Philosophy of Science* 67 (4):580-602.

Shepard, L. A. (1997), "The Centrality of Test Use and Consequences for Test Validity", *Educational Measurement: Issues and Practice* 16 (2).

Sobel, S. E. (1994), "Causal Inference in Latent Variable Models", in A. von Eye and C.C. Clogg (eds.), *Latent Variables Analysis*, Thousand Oaks: Sage.

Spearman, Charles (1904), ""General Intelligence" Objectively Determined and Measured", *American Journal of Psychology* 5:201-293.

——— (1927), *The Abilities of Man: Their Nature and Measurement*. New York: Macmillan.

Staudinger, U. M., and U. Lindenberger, eds. (2003), *Understanding Human Development*. Dordrecht: Kluwer.

Steele, C. M., and J. Aronson (1995), "Stereotype Threat and the Intellectual Test Performance of African-Americans", *Journal of personality and Social Psychology* 69 (5):797–811.

Stoel, R. D., F. G. Garre, C. Dolan, and G. van den Wittenboer (2006), "On the Likelihood Ratio Test in Structural Equation Modeling when Parameters are Subject to Boundary Constraints", *Psychological Methods* 11:439–455.

Thurstone, L. L. (1938), *Primary Mental Abilities*. Vol. 1, *Psychometric Monographs*. Chicago: University of Chicago Press.

Trout, J. D. (1998), *Measuring the Intentional World: Realism, Naturalism, and Quantitative Methods in the behavioral sciences*. Oxford: Oxford University Press.

Undheim, J. O., and J-E. Gustafsson (1987), "The Hierarchical Organization of Cognitive Abilities: Restoring General Intelligence through the Use of Linear Structural Relations (LISREL)", *Multivariate Behavioral Research* 22:149–171.

van der Maas, H. L. J., C. V. Dolan, R. P. P. P. Grasman, J. M. Wicherts, H. M. Huizenga, and M. E. J. Raijmakers (2006), "A Dynamical Model of General Intelligence: the Positive Manifold of Intelligence by Mutualism", *Psychological Review* 113 (4):842–861.

van der Sluis, S., C. Derom, E. Thiery, M. Bartels, T. J. C. Polderman, F. C. Verhulst, N. Jacobs, S. van Gestel, E. J. C. de Geus, C. V. Dolan, D. I. Boomsma, and D. Posthuma (2007), "Sex differences on the WISC-R in Belgium and the Netherlands", *Intelligence* 36:48–67.

van der Sluis, S., D. Posthuma, C.V. Dolan, E. J. C. de Geus, R. Colom, and D. I. Boomsma (2005), "Sex differences on the Dutch WAIS-IIIR", *Intelligence* 34:273–289.

van Fraassen, Bas (1980), *The Scientific Image*. Oxford: Clarendon Press.

Vernon, P.E. (1963), *Personality Assessment*. London: Methuen.

von Eye, A., and C. C. Clogg, eds. (1994), *Latent Variables Analysis*. Thousand Oaks: Sage.

Wechsler, D. (1955), *Wechsler Adult Intelligence Scale*. New York: Psychological Corporation.

Wherry, R. J. (1959), "Hierarchical Factor Solutions without Rotation", *Psychometrika* 24 (1):45–51.

Wicherts, J. M., C. V. Dolan, and D. J. Hessen (2005), "Stereotype Threat and Group Differences in Test Performance: A Question of Measurement Invariance", *Journal of Personality and Social Psychology* 89:696–716.

Wicherts, J. M., C. V. Dolan, D. J. Hessen, P. Oosterveld, G. C. M. van Baal, D. I. Boomsma, and M. M. Span (2004), "Are Intelligence Tests Measurement Invariant over Time? Investigating the Nature of the Flynn Effect", *Intelligence* 32:509–537.

Yung, Y.-F., D. Thissen, and L. McLeod (1999), "On the Relationship between the Higher-order Factor Model and the Hierarchical Factor Model", *Psychometrika* 64 (2):113–128.

Zenderland, Leila (1998), *Measuring Minds: Henry Goddard and the Origins of American Intelligence Testing*. Cambridge: Cambridge University Press.

CURRICULUM VITAE

Steven Brian Hood
sbrianhood@gmail.com

Department of History and Philosophy of Science
1011 E. 3rd St.
Goodbody Hall 130
Bloomington, IN 47405
812-855-3622 (office)
812-855-3631 (fax)

6023 S. Rice Ave.
Bellaire, TX
77401
812-322-6162 (mobile)

EDUCATION

Fall 2002–2008
Indiana University, Bloomington (IUB)
Ph.D. in History and Philosophy of Science
Dissertation Title: *Latent Variable Realism in Psychometrics* (Chair: Colin Allen)

M.A. in History and Philosophy of Science (2005)

2000–2003
University of Florida, Gainesville (UF)
M.A. in Philosophy
Thesis Title: *Theories of Probabilistic Causality* (Chair: Chuang Liu)

1996–2000
University of North Carolina at Greensboro, Greensboro (UNCG)
B.A. *magna cum laude* in Philosophy. Honors Degree
Thesis Title: *The Limitations of Kitcher's Theory of Explanation via Propositional Systematization* (Advisor: Jarrett Leplin)

SPECIALIZATION AND COMPETENCE

AOS: Philosophy of Science, Philosophy of Psychology, Methodology
AOC: Epistemology, Logic, History of Philosophy of Science

ACADEMIC POSITIONS

Bucknell University
Department of Philosophy
2008–2009 Visiting Assistant Professor

Indiana University
Department of Philosophy
Fall 2007 Associate Instructor

Indiana University
Department of History and Philosophy of Science
2003–2007 Associate Instructor

University of Florida
Department of Philosophy
2000–2002 Teaching Assistant

John Hopkins University
Center for Talented Youth
Summers 2001, 2002, 2004, 2005, 2007, 2008 Instructor

PUBLICATIONS

Hood, S. Brian, (forthcoming) "A Comment On Borsboom's Typology of Measurement Theoretic Variables and Michell's Assessment of Psychometrics as "Pathological Science", *Measurement: Interdisciplinary Research and Perspectives*.

Hood, S. Brian, (accepted, revise/resubmit) "Validity in Psychological Testing and Scientific Realism", *Theory and Psychology*.

Hood, S. Brian, (2007) "Book Review: *The Measure of Merit: Talents, Intelligence, and Inequality in the French and American Republics, 1750–1940*, by John Carson, 2007", *Historical Studies in the Physical and Biological Sciences*, vol. 37, no. 2, pp. 480–481.

Submitted for publication: Hood, S. Brian, Dolan, C. V., & Borsboom, D. Working Title: "The Bifactor Model and the Higher Order Factor Model of General Intelligence: Theoretical and Psychometric Considerations"

PRESENTATIONS

"Objections to Michell's Diagnosis of Psychometrics as Pathological Science." Tilburg University, Department of Methodology and Statistics, Tilburg, The Netherlands, March 11, 2008.

"Realist Presuppositions in Psychometrics." Tilburg University, Tilburg Center for Logic and Philosophy of Science, Tilburg, The Netherlands, February 14, 2008.

"Realist Presuppositions in Psychometrics." University of Amsterdam, Department of Psychological Methods, Amsterdam, The Netherlands, November 21, 2007.

"The Anglo-American Tradition in Philosophy." Telavi, Georgia, May 22, 2007 (Sponsored by the US Embassy, Tbilisi, Georgia).

"The Anglo-American Tradition in Philosophy." Gori, Georgia, May 16, 2007 (Sponsored by the US Embassy, Tbilisi, Georgia).

"Philosophy of Science and Psychometric Validity." Indiana University, Inquiry Methodology Program (School of Education), Bloomington, IN, April 24, 2007.

"Validity and Scientific Realism." Indiana University, History and Philosophy of Science, Bloomington, IN, November 17, 2006.

"Leibniz and Verificationism." Midsouth Philosophy Conference, Memphis, TN, February 22–23, 2002.

"Answering Some Objections to Scientific Realism." Florida Philosophical Association, DeLand, FL, November 9, 2001.

AWARDS, FELLOWSHIPS, AND HONORS

Victor E. Thoren Graduate Student Research Fellowship (May, 2008)

Visiting Research Fellow, Tilburg's Center for Logic and Philosophy of Science (Tilburg University, The Netherlands) (February-March, 2008)

Visiting Researcher, Department of Psychological Methods (University of Amsterdam, The Netherlands) (January 2008)

Grant in Aid of Research (IUB) (2007)

S. Westfall Fellowship for Graduate Student Research Travel (IUB) (awarded twice) (Spring, Fall 2007)

Ruth N. Halls Fellowship (IUB)

Social Science Research Council Grant for Language Study (IUB) (Summer 2006)

Phi Beta Kappa

Josephine Hege Award (*Phi Beta Kappa*) (UNCG) (2000)

Bernice Love Stadiem Memorial Scholarship (UNCG)

Phi Sigma Tau (International Honors Fraternity in Philosophy) (UNCG)

Fred C. Koch Scholarship (awarded twice) (1996, 1997)

Dean's List UNC-Greensboro (1997–2000)

TEACHING EXPERIENCE

Instructor: responsibilities include lecturing, grading, course design

Intelligence for Everyone (History and Philosophy of Psychometrics)
Fall 2005, Spring 2006, and Fall 2006 (IUB)

Scientific Reasoning
Fall 2004, Spring 2005, and Spring 2006 (IUB)

Introduction to Philosophy
Summer 2005 (including the supervision of a teaching assistant) (Johns Hopkins University Center for Talented Youth (CTY)), Bristol, RI

Logic: Principles of Reasoning
Summer 2001, 2002, 2007 (including the supervision of a teaching assistant) (CTY), Lancaster, PA

Introduction to Logic
Summer 2004, 2005 (including the supervision of a teaching assistant) (CTY), Bristol, RI

LSAT Workshop
Spring 2004 (IUB-Prelaw Center)

Teaching Assistant: responsibilities include grading, leading discussion sections

Introduction to Philosophy (worked with Fred Schmitt)
Fall 2007 (IUB)

Genetics, Eugenics, and Biotechnology (worked with Sander Gliboff)
Spring 2004 (IUB)

Quantum Mysteries for Everyone (worked with Michael Dickson)
Fall 2003 (IUB)

Philosophical Writing (worked with Greg Ray and John Palmer)
Fall 2001 and Spring 2002 (UF)
Introduction to Logic (worked with Robert Baum and Marin Smilov)
Fall 2000 and Spring 2001 (UF)

RESEARCH ASSISTANTSHIPS

Elisabeth Lloyd, Summer 2005 (IUB, History and Philosophy of Science)
Michael Dickson Summer 2004 (IUB, History and Philosophy of Science)
Jarrett Leplin 2000 (UNCG, Philosophy)

SERVICE

Presenter for the US Embassy in Tbilisi Georgia's American Corners
Co-director for Indiana University's "Human Intelligence" website:
http://www.indiana.edu/~intell/index.shtml
Graduate student representative (IUB)
Vice President of the Graduate Student Philosophical Society, 2001-2002 (UF)
Co-organizer of the University of Florida-Florida State University Graduate Student Philosophy conference (UF)
University Marshal (UNCG)
Environmental Awareness Foundation: Student Government Representative (UNCG)

PROFESSIONAL AFFILIATIONS

American Philosophical Association
Psychometric Society

LANGUAGES

English: fluent native speaker
Russian: reading knowledge only
Georgian: reading knowledge and elementary conversation
Spanish: reading knowledge and elementary conversation